International Journal of Computer Networks and Communications Security

VOL.1, NO.6, NOVEMBER 2013, 237–250
Available online at: www.ijcncs.org
ISSN 2308-9830

Data Privacy: An End-User Perspective

Esma Aïmeur[1], Gilles Brassard[2], Jonathan Rioux[3]

[1, 2, 3] Département d'informatique et de recherche opérationnelle, Université de Montréal

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

ABSTRACT

Online privacy has become a raising concern for digital citizens within the past few years. As users of the Internet, we are in an uncomfortable situation regarding the protection of our online data. We share, tweet, like and follow at an ever-increasing rate, while at the same time getting more aware of the possible dangers of privacy breach or identity theft. Is it possible to navigate on the web while being sure we're not being spied on? This paper highlights the main data collection fields, taking a user's view on the risks and tradeoffs regarding online data collection and privacy.

Keywords: *Privacy, User Experience, Data Collection, Social Media, Identity Theft.*

1 INTRODUCTION

The Internet has quickly become a central part of our lives. We now spend hours browsing, writing emails, frequenting social media and sharing with loved ones and people all over the globe. We made great progress bringing more and more people online and the web isn't a haven for scientists anymore.

We're now witnessing a democratization of online access, combined with a focus on web technologies. Websites are trying to tailor themselves to their customers, gathering and using the information they are providing in order to offer a differentiated product. Most people are aware of their browser's history and cookies, but with the rise of single-login, geolocation and online profiles, the boundaries are getting blurrier. Companies are collecting data at an exponential rate. For example, Facebook is investing massively in new technologies to deal with never seen before amounts of data: Apache *Cassandra*[1] was developed by one of their team and was scaled to over 300 TB of data. More recently, they open-sourced *Presto*[2], which can scan over a petabyte of data a day.

With the rise of mobile phones and tablets, "seamless" is now much more than a buzzword: service providers are now competing to provide the most convenient, most transparent experience for the user. Our data is now stored in the *cloud* and it's not only normal to access it anywhere, it is desirable. The web is now *social:* blogging is now more accessible than ever, with Facebook, Twitter, Instagram, Pinterest and many other knowing their golden era. It is now obvious that we're living in a digital economy considering how the content creation is now driven by the users (we can talk about *user-generated content*). Many companies on the World Wide Web are now worth millions, even billions of dollars, with nothing more for assets than their huge data warehouses and their analytic toolkit.

This is perplexing: how and when did our social and personal data become so interesting, so profitable? What exactly are *they* collecting and why? When we're posting, tweeting, sharing posts or articles online, who is able to access that contents and what can *they* gain from such insight? While sharing information brings many advantages, brought to our attention by the different companies, it is obvious that they're gaining a lot from us sharing more and more online. Approximately 60% of all Internet end devices exchange traffic with Google servers during the course of an average day [1], Facebook now has whooping 1+ billion users online and the US government was recently blamed over the NSA recent PRISM program scandal; it is safe to assume that they know a lot more about us

[1] http://cassandra.apache.org/
[2] http://prestodb.io/

than anybody. (As the joke goes, "the NSA is the only governmental institution that listens to the people"… but how do we wish it were only a joke!) This can be pretty unsettling and we have good reasons to try to keep our identity safe, away from unwanted data collectors or identity theft. The question is: *are we able to*?

2 INTERNET DATA COLLECTION

We are collecting and providing data from the moment our computer is connected to the Internet. It is important to remember that most data aggregation techniques are advertised as a mean to ease out the web experience for the web user. Targeting a particular customer on his[3] preferences, avoiding duplicate information during a web search, ensuring a smoother experience from cross-browsing by keeping handy all the login information to one's favourite websites; the examples are endless.

The next few sections explore various domains in which Internet data aggregation and analysis are used, as well as the primary uses, challenges and threats.

2.1 Healthcare

Information technologies (IT) can be very useful for the storage of patient records or the execution of repetitive tasks in a clinical environment, thereby opening the door to potential cost reduction and process optimization. Examples of possible Healthcare Information Technology (HIT) contributions include assistance in diagnoses and prescriptions, automatic dosage of medication, easier access to a patient's medical history, management of patient schedules and priorities, leading to a reduction of waiting and staying times. A case study of the changes engendered by the adoption of IT in Kyoto, Japan, showed data demonstrating the positive impacts of IT in a hospital environment [2]. Basically, it has been demonstrated that the introduction of automated medical systems could improve the relationship between the patients and the physicians, compensate for staff shortages, ease the jobs of nurses (thus promoting recruitment and retention) and encourage collaboration between distant health facilities.

As an example, the US Congress' American Recovery and Reinvestment Act of 2009 allowed a budget of $20 billion for healthcare information technology [3]. Similarly, Canada has been digitizing all its citizen's medical records since 2012, implementing an EHR (Electronic Health Record) at an estimated cost of $10 billion. The recorded information ranges from an individual's features to his medial data. This includes traits such as height, waist, body fat, but more importantly, past diseases, last visits to emergency clinics, fertility status, emotional problems, lifestyle, and more [4].

Needless to say, health data is sensitive and should not be prone to bugs, unintended uses or security flaws. Moreover, one major concern regarding the centralization of data as sensitive as health information is the protection of privacy. What if a technologically knowledgeable person were able to hack into EHR data servers and steal all the information it contains? One could sell this information to pharmaceutical companies, steal someone's identity or malevolently use the information to exploit one's medical weaknesses.

Although there are definite advantages and benefits for health service users, it is important to ensure that EHRs are used in a way that the integrity of personal information is preserved and that patients have control and access to their records. A debate that has yet to be resolved is whether a patient can have unrestricted access to his own records. Or more specifically, who owns this information? There are many diverging points of view on this issue, but it is still difficult to find a balance between the well being of a patient and the preservation of his privacy.

2.2 Online-based Companies

Marketing and advertising companies have understood the power of information for a very long time. The more they know about demographics, consumer habits and preferences of particular customer types, the more they can tailor their product offerings, and as a result, the more sales they can make [5]. Before the Internet, most of the information was gathered with the shopper *knowingly* using a fidelity card (*e.g.* Air Miles). As the consumer is steadily moving online, we are also seeing a shift in how data is collected and used.

Technology also has a say in the field of reputation preservation. For example, the notion of social shopping is gaining grounds in today's purchasing processes. People give their opinions on products and companies on social networks, blogs and forums and uncontrollable information is circulating on the Internet. Up to 55% of online shoppers buy products from businesses about which they have knowledge and 38% are converted into

[3] "His" refers to "his or her"; "He" refers to "he or she", when applicable.

customers because they heard about a brand from another person. This explains why companies need to put effort into making word of mouth as positive as possible.

Additionally, mobile payment technology is spreading throughout consumers as a new convenient standard for purchasing. *Square*[4] is a rising star in this world, making it possible for small businesses to accept credit cards with minimal investment. Their new products allow users to "check-in" with their phone, place an order, and the merchant simply has to tap on your picture from his terminal to complete the transaction. Technology has truly the potential to disrupt consumers' habits while buying physical or online goods. On the other hand, businesses will not only face some pressure not only to keep in touch with the technical evolution of this payment method, they will also need to be aware of the potential security threats it incurs. Hackers are always one single step behind the development of a new technology, and if mobile payments have vulnerabilities, companies should be aware of them as much as cybercriminals, if not more.

Finally, we're currently seeing the booming of the digital economy, Twitter and Facebook's IPO being one of the notable examples. Those data-driven companies are now worth billions of dollars with not much assets beside their consumer base. Those major players are growing very quickly, acquiring start-ups and existing companies (as well as their data) to consolidate their position. The valuation of those companies is often based on their capacity to monetize this consumer base into advertising profits or sales. It appears obvious that our personal and consumer information is something valuable in the eyes of advertisers and sellers alike.

2.3 E-Learning

One of the main advantages of E-learning and Intelligent Tutoring Systems is their adaptability to the learner's specific needs and preferences. Nonetheless, to do so, these systems collect large amounts of information about the learner. This information could be misused, thus violating the learner's *privacy*, which is the right of individuals to determine what information about themselves is known to others, as well as when and how it is used [6].

Although the security of E-learning systems is imperative to preserve privacy, it is not sufficient. Indeed, security will protect learners' information against unwarranted access, but not against abuse from authorized access. Enforcing the Integrity and Confidentiality of the learner's information does protect the learner's data (and consequently his privacy) from unauthorized access. Indeed, E-learning systems gather large amounts of information about the learners—information that is readily made available for the tutors, but also for E-learning platform system administrators [7].

Specifically, privacy is nearly absent in current E-learning systems. Only primitive forms of privacy are offered in some platforms. For instance, restricting tutor access to certain pieces of data such as auto-evaluations performed by the learners. Nonetheless, the tutor has access to virtually all the remaining information including, but not limited to, who the students are, what parts of the course they referred to, how many times and for how long, as well as all the messages in the forums and all the information about the quizzes and tests the learner took in his course. As a matter of fact, the *DataLossDB*[5] (data loss database) regularly reports on breaches concerning student records.

2.4 Personal Online Data Storage

Online Data Storage can be considered to be the more modern and convenient offspring of the USB Key. Many solutions now exist, such as *Dropbox, Box.com, SpiderOak*, Google *Drive*, Microsoft *SkyDrive* and so on. Most of the time, those solutions advertise themselves to be a hassle-free, seamless synchronization experience.

As those platforms prosper on the web (Dropbox reached the 175 million-user mark during the summer of 2013, Microsoft SkyDrive has now more than 250 million users), a growing concern about the safety of personal data stored "in the cloud" is emerging. The website *dumpdropbox.com* addresses some fears about this popular platform in a rather concise way.

Privacy is, of course, a growing concern for those providers: we, as users of one or many online storage services, are willingly uploading our files to their servers and we are relying on their privacy-enabling technologies to keep our data safe. As we are more aware of what we're uploading, most users are more careful about how their data is stored. To address this concern, every provider has a page explaining how their storage is working. Some solutions are even using this as a differentiation factor: SpiderOak boasts on their main page a "Zero-Knowledge privacy environment" and claims that nobody, even their employees, can access a user's data. Dropbox also

[4] https://squareup.com/ca

[5] http://datalossdb.org/

has a wordy security policy stating that employees can only access the file's metadata [8].

3 INTERNET DATA COLLECTION TECHNIQUES AND CONSEQUENCES

According to Schneier [9], "We leave data everywhere we go. It's not just our bank accounts and stock portfolios, or our itemized bills, listing every credit card purchase and telephone call we make… It's also our lives."

3.1 Social Media

According to Experian Marketing Services [10], 27% of the time spent online is on social networking, which is more than any other online activity. By their very nature, those websites aggregate, classify and collect various data about our preferences (*Likes, Shares, (Re-) tweets,* etc.), our opinions, what we follow. As they try to mimic our day-to-day life, social networks can provide to marketers and companies better insight about how we shop, how we judge products and services and how we share our preferences [5]. Every social media platform provides a streamlined mean of advertising on their websites, often targeting a subset of the general population to maximize efficiency. This is a direct contract with the classic, undifferentiated way of doing online advertising, where you try to put your brand wherever you can. This is a win-win situation for the consumer and the advertiser: the consumer can be presented with products and services more in tune with his own liking and the advertiser can be promised better conversion (click-to-pay) rates.

Even without trying to gain illegal access to private data, we can easily gather relevant information by browsing the public information showcased by individuals. Privacy policies are often tedious to read and they changes on a regular basis, making the user often click on the "accept" button without reading. According to McDonald and Cranor [11], the average privacy policy can take 10 minutes to read (approx. 2500 words), an unacceptable length of time for most people. Knowing what we share and with whom we're sharing it is an "always on-going" process and not being perpetually vigilant erases our efforts.

As we'll see later, our social profile contains very interesting information for data collectors, may they be insurance companies, background check firms or online pirates. Cross-validating secret questions, often defaulted to easy-to-learn, predictable questions, can be a serious threat to a careless user.

3.2 Online Data Brokers

A debatable business model that followed the evolution of technology is the world of online data brokers. There are websites such as *Abika.com* or *USSearch.com* that, for a fee (sometimes for free), let anyone search for a name in order to retrieve all the personal information about him that is available in a multitude of public records. Possible data include the person's name, address, date of birth, marital status, age of children, list of relatives, mortgage information, bankruptcy history and even sensitive information such as Social Security Numbers, voting records or court records [12].

As an example taken from a 2006 news article, Ed Whitfield, a Republican militating against this kind of business, stipulated that one could buy someone's cell phone record for $200, Social Security information for $60 and a student's university class schedule for $80. Needless to say, several privacy principles are violated by this type of service.

3.3 Search Engines

Search tools such as *123people.com, Whozat.com, Pipl.com, Peekyou.com, PeopleSearch.net, Peoplefinder.com, AnyWho, Yasni.com,* are also good sources of information for administrators. They are free real-time people search tools that look into nearly every corner of the web to provide and gather information. There are also social network aggregator web sites such as *Lifehacker.com, Spokeo.com, Spoke.com* and *Intelius.com*, which collect data from various, online and offline sources (phone directories, social networks, etc.) and have large databases from which they may unknowingly sell information to malicious people [13].

Moreover, when one searches for a given product, there are many variations on how often he may search for it (searches for "chocolate" predictably increase before Christmas). Thus, whenever there is a sudden spike in the number of Google queries for a given term, it probably indicates that something extraordinary has just happened; the likelihood is even higher if the search spike is limited to a particular geographic area. For example, when an unusually high number of Internet users in Mexico began Googling terms like "flu" and "cold" in mid-April 2009, it signalled the outbreak of swine flu [5].

3.4 Geolocation

Most of today's mobile phones are equipped with a Global Positioning System (GPS) chip, allowing people to know where they are located at any instant. Not only does the GPS user have access to his location, so do applications residing on the device (perhaps after asking permission from the owner). This is now known as geolocation. Aside from the well-known map functionality made possible by this technology, there are many interesting applications such as FourSquare*,* *FacebookPlaces* and *Gowalla*. They are used to indicate your location to friends and, conversely, see their location. Similarly, Twitter has the option of attaching the user's location to its tweets. In the absence of better alternatives, the features of geolocation can be used to track the location of a vehicle that belongs to a company.

Another popular product using geolocation is mobile online dating. A quick research on Google Play or the App Store gives many results, mostly targeted to the homosexual community. Those apps go beyond the regular dating website by ordering the results according to your position and can go as far as telling you the distance from where you are currently located.

Furthermore, Geographic Information Retrieval (GIR) systems are increasing in popularity. These systems can capture user needs from their queries by processing them, matching them against the users past geographic tendencies. Such information can be shared between different organizations through the GeoNetwork open source project [14].

3.5 Background Check

As we're adding content each and every day on social networks, writing in blogs, and commenting on websites, we are often unaware on how much of this information is freely available for anybody to see. This provides a tremendous source of information for future employers and background check firms [15]. Since a company wants to minimize hiring risks, it will certainly refrain from hiring someone having tasteless pictures of him wandering online or expressing dangerous opinions whenever he gets the chance.

Cleaning up our social presence can prove harder than it seems. Given the rate at which we share, publish, like or tweet contents, it can prove hard to screen our own personal history on multiple websites : googling ourselves is often not enough. It is unfortunately safe to assume that our online life is as permanent as a tattoo [16] and that, given enough time, anything can be found out.

The insurance industry also watches the Internet very carefully: property and casualty insurance are investigating if disclosing the fact that you're away from home for an extended period of time on a public website (such as Facebook) can fall into a breach of the "duty of care". Health insurers can also take the information opportunity to ensure that a disabled patient isn't using the claim payments and the time off to take some vacations. Many examples of such cases have been featured in the media during the past few years.

3.6 Skype and Online Conversations

Online conversations are clearly one of the main uses of the Internet: the rise of many webchat applications in the past few years (*Facebook Chat*, Google *Hangouts*, etc.) is a clear sign of this. IRC, even though we hear less and less about it, which is still very active (approximately 400 000 simultaneously connected users on the top 100 servers at the time being [17]), continues to be one of the most popular decentralized platforms. Since its creation, the use of SSL (secure socket layer) was slowly introduced for client-to-server, thereby increasing the difficulty for malevolent people to eavesdrop on a conversation.

Skype is also a major player in the field. Since its merge with Microsoft Live Messenger, one of the most popular Internet Messaging (IM) platforms, Skype is now ubiquitous (approximately 10% of the world population has a Skype account [18]) to many people who are using it to keep in touch via its tremendous and well-known video chat ability.

On the other hand, Skype doesn't have an enviable reputation when it comes to data privacy. It is now well known that every conversation is taped and recorded [18, 19]. Its privacy policy is clear on this.

"Skype may use automated scanning within Instant Messages and SMS to (a) identify suspected spam and/or (b) identify URLs that have been previously flagged as spam, fraud, or phishing links. In limited instances, Skype may capture and manually review instant messages or SMS in connection with Spam prevention efforts."

The software company got into trouble a few notable times before and since its merger: last march, a student of the University of New Mexico was able to prove the surveillance and censorship capacities of a modified version of Skype targeting the Chinese population: TOM-Skype [20]. With the recent NSA scandal and Microsoft's position with the US government, Skype is clearly a high-profile player and its poor privacy reputation, combined with its massive popularity, is something to be worried about.

3.7 Mobile Phones and Applications

As 70% of the Canadian market is using a mobile device of some kind, the Internet landscape drastically changed within the past few years. More and more websites are now tailored to their mobile users and the notion of "always connected" is now a reality.

Two main operating systems are now shaping the mobile market: Apple's iOS and Google's Android. Emerging from these two platforms, hundreds of thousands of applications, or "apps", are available from Google Play or Apple's AppStore. Downloading an app usually implies accepting a privacy policy, where the app maker asks for permission(s) to access certain components of his phone. As some applications are tightly coupled with the users' personal data, some can go as far as requiring full access to the SMS information or the users' own address book. Both platforms express guidelines about how and what permissions should be asked. Unfortunately, this does not prevent abuses from some app designers and the end users must be vigilant at all time when downloading or upgrading an app.

Most mobiles platforms use two-way synchronization in order to keep the users' information up-to-date. This requires a considerable amount of trust from the user base, since the information shared can be very intimate and diverse: contacts, emails, passwords, credit card number, browser history, WiFi hotspot WPA keys, etc. The mobile-to-computer experience is narrowing the gap, allowing users to share all the information collected from one browser to another. Google Chrome makes this its default behaviour. This is advertised as an "easier, friendlier browser". It is debatable if a web browser should offer the synchronization of personal and sensitive information by default or if it should be explicitly requested by the end user.

Besides sharing one's own personal information, most social applications (Viber, Whatsapp, etc.) are asking their users to get access to their contacts. A third party can therefore identify you not only when you register to their service, but also when someone who has you as a contact uses it. Most of the time, contact entries contain the full name, phone number, email and, more rarely, full address.

The size of these devices and their monetary value also is a data-security hazard as they're easily stolen. An improper lock mechanism combined with the lack of encryption on the device can give access to the owner's full profile in a matter of minutes. As we're shifting from computers to mobile devices for casual and business browsing on-the-go, this is a new and rapidly growing consideration that we can't ignore.

3.8 Identity Theft: The Ultimate Consequence of Information Disclosure

It should be obvious that people want to protect information that makes them vulnerable. That is, information that can be used by others to threaten them physically, emotionally, financially or harm their reputation [21].

Solove [21] describes privacy violation according to four categories: Information Collection, Information Processing, Information Dissemination and Intrusion. The possible threats are many: surveillance, interrogation, aggregation, identification, insecurity, secondary use, exclusion, breach of confidentiality, disclosure, exposure, increased accessibility, blackmail, appropriation, distortion, etc.

Identity theft is one of the most dangerous among these—the ultimate consequence of Internet data collection. When people surf the Web, make purchases, bank online, communicate via email or instant messaging, or even visit gaming sites on the Internet, they are regularly exposed to major risks including the violation of their privacy [13]. So, identity theft is made possible because we all have "digital dossiers"—extensive repositories of personal information about us—that are maintained by various companies and institutions [22].

It should be noted that it is difficult to measure a prejudice when it is not financial. Victims of identity theft are submerged into a bureaucratic nightmare, having to spend nearly two years and almost 200 hours in order to decontaminate their dossier [22]. During this long process, victims encounter difficulties getting jobs, loans or mortgages. Combining this with the growing ubiquity of online banking and services make identify theft increasingly damageable.

4 PRIVACY ENHANCED TECHNOLOGIES (PETs)

While we can often fine grain what we share with various platforms, some users prefer to keep the control on how privately they surf the Internet. In order to ensure privacy and anonymity online, some technologies have been made available to the public. Shen and Pearson [23] conducted a survey of Privacy Enhancing Technologies (PETs); we summarize below the most important types of PETs according to their paper.

4.1 PETs for Anonymization

One important technology for preserving privacy is anonymity [24, 25, 26], which implements data minimization and user identity protection techniques, aimed at preserving privacy at different levels. For anonymous Communication Techniques, various technologies such as Hordes [27], Crowds [24], Anonymizer1, and private authentication protocols for mobile scenarios [28], have been proposed to keep users anonymous. Dingledine, Mathewson and Syverson [26] have introduced Tor, a well-known circuit-based low-latency anonymous communication service, which addresses perfect forward secrecy, congestion control, directory servers, integrity checking, configurable exit policies, and a practical design for location-hidden services via rendezvous points. Some ISP and servers won't allow Tor for the time being.

4.2 PETs for Identity Management

Identity management deals with identifying individuals and controlling access to resources in a system. There are several approaches in this area. In particular, Liberty Alliance's federated approach7, OpenID8 authentication (a decentralized approach), Identity Metasystem Architecture [29] and Generic Bootstrapping Architecture (GBA) (focused telecommunication). More specifically, there are credential systems that allow authentication (and authorization) without identification by providing only the Personally Identifiable Information (PII) necessary for the transaction or a proof of entitlement. Among these credential systems, one may mention those proposed by Chaum [30] and Brands [31], as well as credential identity management [30, 32, 33].

4.3 PETs for Data Processing

In the field of data mining, various methods have been proposed to minimize access to users' private data: additive data perturbation [34, 35], multiplicative data perturbation [36, 37], data anonymization [38, 39, 40], secure multi-party computation [41, 42], privacy-preserving multivariate statistical analysis [43, 44], probabilistic automata [45], privacy-preserving formal methods [46], sampling-based methods [47], *k*-anonymization classification [48], privacy in graph data [49], statistical disclosure control [50], etc.

5 RE-IDENTIFICATION

Despite all the tools cited above, it is possible to re-identify people—that is, to determine the exact identity of a person by gathering and linking various pieces of information disseminated across the web. This process has recently gained in popularity following the emergence of personal databases on the Internet. As a consequence, there is no absolute protection, even with Privacy Enhancing Technologies.

5.1 Linkage

The most widely used method is linking, or "data linkage", which combines information from a separate entity to learn more about an individual. Initially, one can merge two data sets (often databases) to form one. Repeating this process iteratively increases the amount of useful information. Latanya Sweeney was the first to demonstrate the effectiveness of linkage when she managed to find the Governor of Massachusetts in an anonymous medical database. She simply linked this medical information to a database of voters in Cambridge by identifying common information between the two datasets (zip code, date of birth, sex) [40].

5.2 Inference

The process of deductive inference, as its name implies, consists of using basic information to deduce new information (e.g. guess the person's tastes from their habits). In 2009, a group of researchers showed it was possible to figure out an American's social security number by knowing the person's date of birth and geographical location [51]. Note that these two pieces of information, which are sufficient to trace the victim's social security number, are very easy to obtain (mainly on the Internet). However, social security information is very sensitive and needs to be protected, as it can allow criminals to steal someone's identity. Inferences can be made from collected data, as was the case for the study "On the Anonymity of Home/Work", in which two researchers were able to spot the residence and workplace locations of Americans, simply from movements acquired through their GPS device [52].

5.3 Homogeneity

Re-identification by homogeneity consists of linking a subject's traits, belongings or habits to a homogeneous group in order to deduce all, or part of his identity [38]. Consider a database in which anonymous users are classified into groups according to their age. Finding the age of a given user determines the group to which he belongs, therefore reducing the range of possibilities regarding the person's identity. If extra steps are added to this process, for example adding a gender filter to the group, one can eventually end up inferring the exact identity of the individual.

5.4 Graphs and Machine Learning

More recent sophisticated techniques, based on machine learning and the linking of graph nodes, are emerging. In 2009, two researchers at the University of Texas at Austin developed a generic algorithm for re-identifying anonymous social networks [53]. This algorithm uses the structure of social networks—that is, relations between users (e.g. the "follow" relationships on Twitter, "friends" on Facebook). Linking between the nodes of two graphs can group user profiles together. The results are then verified by analysing the correspondence between the usernames or names shown in the different profiles.

One common factor that emerges from all these techniques is the almost systematic usage of information that does not identify an individual directly, but rather by combining various pieces of data through a thorough search. It is then possible to successfully identify the individual. This concept was defined for the first time in 1986 by Dalenius, who named it "quasi-identifier" [54].

6 DISCUSSION

6.1 Problems with privacy

One of the most important problems is that despite all the criticism one could direct towards collectors, most of the time people are actually informed of what type of data is being collected about them and how it will be used, often by the means of a privacy policy. One could then argue that users should be able to decide whether or not they accept such collection and usage, and it is their responsibility to make the choice that is right for them [21]. But even for those (few) people who take the time to read privacy policies, they often lack the expertise to adequately assess the consequences of agreeing to the collection, usage or disclosure of their personal data.

But maybe this would not be such an issue if the websites that perform the act of collection weren't so numerous. Even if each company provided an easy and clear way to handle privacy issues, there would still be too many of them. The average person visits nearly a hundred websites per month, doing business online and offline with countless companies (merchants, utilities, insurance, technology, travel, financial, etc.). The point is that it is really difficult for an individual to manage a hundred privacy mechanics at the same time, as clear as the policies may be. And even if it were the case, people's lives are not a set of fixed points in time, and the contexts change. If people's preferences are fluid, they might consent to the collection, usage or disclosure of their data in one situation but change their preferences in a different context.

As if this weren't enough, there is also an issue of perception about the possible usages and consequences of collected data. Instinctively, an individual giving out an innocuous piece of personal information is likely to think he is not revealing anything sensitive. Thus, at other points in time, the same person may reveal equally non-sensitive data, and soon enough there is sufficient information scattered in various places to combine it, analyse it and obtain sensitive facts about the person. This would never have been anticipated; as such derived facts were never given away in the first place. However, this art of re-identification is similar to a detective's work, and the more clues you leave behind, the more likely you are to be traced.

It should also be said that consumers have minimal bargaining power regarding their privacy in various commercial transactions; often access to purchases is under the condition of policy acceptance. One can withdraw from such consent, but this has to be specified explicitly to the business. Some argue that a better solution is to move to an opt-in rather than opt-out regime, such as in Europe, where affirmative opt-in consent is often required. As stated by FTC Commissioner Jon Leibowitz, companies should move to a model in which consumers "opt-in when it comes to collecting information—especially when it comes to sharing consumer information with third parties and sharing it across various web-based services." [55]. It can also prove very hard to clean up one's Internet presence, as some websites won't delete your data when you opt-out of the service or close your account. The website justdelete.me keeps an up-to-date list of the most popular sign-on websites: the existence of such a service is a concern about the amount of control we have as individuals on our own privacy and personal information.

While many Countries now have privacy rules or laws (for instance, Canada has a Privacy Act, overseen by the Privacy Commissioner of Canada[6], while the European Union has rules available on the European Commission website[7]), the web is a global machine. Using a foreign website means having at least two sets of rules and laws with which to abide. While laws may provide a framework to rule and enhance privacy for web users, possibly through the use of PETs, they are indeed powerless about what we *willingly* share (and might regret later). Legal boundaries and technological tools are no substitute for sound and careful practice, since a notable part of the information easily available on the web was consciously placed there: we can't put the blame on a security breach for an inappropriate tweet!

The multiple fusions and acquisitions finally add a layer of complexity for the end user trying to know where his information is located. The buyer might not have the same view on user privacy and removing ourselves efficiently from a website can prove next to impossible as we saw earlier. Since we can't predict where the company behind the services we're using right now will be in the near future, it is harder to keep track efficiently of our personal information.

6.2 Public versus Private

Now, let us picture a world in which all the aforementioned issues have been addressed—a world in which every policy is crystal-clear, users have the means to decide whether or not it is in their best interest to reveal information, and no one is forced to commit their privacy in order to make purchases or gain access to contents. The problem that arises next is that even if people were completely aware of the consequences of personal data disclosure, most of them would still want to publish and share, not caring really about exposing themselves!

In fact, one could argue that this privacy preservation debate is futile, as collecting organizations are not solely responsible for the creation of today's ocean of circulating personal information datasets—they merely built the tools to use it, and it is the users that have been filling it for years. It could be said that the very people privacy advocates are trying to protect are the ones that caused the problem in the first place! And maybe that is the way people actually want it. As David Weinberger puts it: "In the digital world, everything has its places, with transformative effects: Information is now a social asset and should be made public, for anyone to link, organize, and make more valuable; […] More information gives people the hooks to find what they need; Messiness is a digital virtue, leading to new ideas, efficiency, and social knowledge." [56].

But then again, this kind of argument implies that humans always reason adequately before typing on a web browser. Researchers [44] have investigated regrets associated with users' posts on Facebook, which revolved around sensitive topics, contents with strong sentiment, lies and secrets. They found possible causes for why users make posts that they later regret: they do not think about their reason for posting or the consequences of their posts; they misjudge the culture and norms within their social circles; they are in a "hot" state of high emotion when posting, or under the influence of drugs or alcohol; their postings are seen by an unintended audience; they do not foresee how their posts could be perceived by people within their intended audience; and they misunderstand or misuse the Facebook platform. So, in the end, privacy protection might not be about making policies clearer or building the right tools, but it is rather a problem that resides deep in the complex human nature.

Finally, what we think is private might not be: the US National Security Agency's PRISM program was recently uncovered, showing to the public the amplitude of the operation. Microsoft, Google, Yahoo, AOL, Apple, Skype, PalTalk, Facebook and Youtube users got their information logged, according to an investigation from the Washington Post [57]. Although the NSA's original mission is "[…] to protect our nation [the United States] from a wide variety of threats", many people are concerned about the loss of confidence concerning U.S. cloud companies: the Cloud Security Alliance estimated the economic cost to be approximately 35 billion dollars, as 56% of non-US residents have become less likely to use US-based providers in light of the recent PRISM-related events [58]. This is a critical hit for such a volatile and growing industry.

6.3 Caching and Archiving: A Collective Memory of the Internet

Since a webpage isn't something tangible, keeping a history of the Internet isn't as straightforward as filing a book on shelves: we have to consider that a webpage or an electronic document:

[6] http://www.priv.gc.ca/index_e.asp

[7] http://ec.europa.eu/justice/data-protection/

Can disappear once deleted by its author
Can be modified without warning
Can be unavailable for a period of time

Some websites have made the collection, storage and retrieval of electronic documents their business. Google Cache (available from the "Google cache browser" [59], for instance) allows, as its name indicates, end-users to retrieve a cached version of a webpage or document. This can prove tremendously useful when a page is brought down following an attack (mostly DOS or DDOS) or a sudden burst in page views (often called the Slashdot / Reddit / Digg / Hacker News effect [60]). Although a Webmaster can request that the cached version be removed from Google's results, the procedure isn't straightforward and doesn't apply for every kind of documents [61]. Furthermore, the old version will only be removed from the public eye once Google re-indexes the webpage.

The Internet archive (and its most famous product, the Wayback Machine [62]) provides an easy way to access various historic snapshots of a website. Although it can't track each revision of every single webpage, the tool boasts as of today "364 billion web pages saved over time". In order for a website to be "indexed" by the Wayback Machine, it simply has to allow crawler robots.

Those two tools have an obvious role in the preservation of what we can call "the collective memory of the internet": tools like Google Cache provide a short-term memory retrieval from pages that have been brought down for one reason or another while the Internet Archive makes possible the consultation of long-down or older versions of a web document. An archive service brings a way to properly quote or access a specific version of something we consulted previously, considering the three characteristics of online documents previously mentioned

Knowing how those systems works, the privacy issues and concerns are obvious. If sensible information is released online, we have not only to delete the original document, but also every cached or archived copy available. This is an impossible task, as we can't check every crawling agent and remove every local copy some users might have on their hard-drive. Once a document has been published online, we can't simply remove it and be confident that it has completely disappeared.

The procedure to prevent being archived is relatively painless for the Wayback Machine, requiring only the inclusion of a robots.txt file at the top level of the website [63]. Although we salute this simple privacy-enhancing option, we can still raise some concerns about the posterity of online documents: is archiving the same thing as collecting data? Should we keep a copy of every valuable piece of information ever published on the Internet, like a "central library"? How can we filter relevant information from "noise"?

6.4 The Right to Erasure

Another hot topic is the right to remove one's online presence and how the different services (search engines, social networks, etc.) should behave when someone requests self-oblivion. Currently, it is very hand to sanitize one's online presence as many services are keeping an artificial presence, Facebook being the main culprit here. Since anybody is able to tag pictures with anyone (even if the tag isn't linked to a precise profile, it can nevertheless be used for facial recognition), a bad or compromising picture can easily be retrieved.

The European Union recently ruled in favour of Google by vetoing the right to be forgotten [64]. They consider that the removal of "legitimate and legal information" would be the equivalent of censorship.

Another debate that is currently raging concerns the rights of deceased people. How can and should an online platform, especially a social one, deal with the death of one of their users? Slowly, some people are starting to include guidance about such subjects in their wills [65]. Simply put, the death of a person doesn't mean the elimination of his presence from the web and this can lead to a variety of awkward situations.

7 CONCLUSION

Being on the Internet implies constantly sharing information, may it be personal or not. While there are means to limit or acknowledge how much and what we're sharing, many agrees that the current situation is unbearable. To counter this phenomenon, there are various privacy enhancing technologies that may be used, but they will never be sufficient because re-identification is always looming. We're facing a unique, uncomfortable situation: as social media is booming and more and more people are using the web to share information, privacy issues are becoming more complicated, yet increasingly important.

The digital economy is changing at an ever-increasing pace. Being connected is now fundamental for many individuals, and companies are tapping into that market: Google is now trying to change the Internet providers' market by launching Google Fiber[8] in selected cities, an obvious move considering the nature of their business. The more often people are online, the better the outcomes for those digital conglomerates.

We know that collecting, aggregating and using data is the backbone of the Internet: search engines, banking websites, credit rating agencies, caching and archiving services, those platforms are more data-driven than ever before. Every person connected on the Internet takes the role of the end-user one time or another.

Is it worthwhile to try to keep decent privacy online? “Digital Natives”, as they are called by Prensky [66], do not really care about disseminating their information. Even if they are wary of the consequences, they claim that life is for sharing! “They want to be the targets of marketing. They want their data shared. They want to get catalogues mailed to their homes. They want to be tracked. They want to be profiled.” [21] Is it because we are not sufficiently aware of the implications? Is it because the advantages outweigh the inconveniences?

Although an imperfect analogy, we consider that thinking in terms of a “Digital Wallet” that would contain our private information is a powerful image to convey the importance of privacy. As a real wallet contains (beside cash) precious information that could be dangerous in the wrong hands, a digital wallet, improperly secured, can lead to undesirable consequences to one’s online experience.

Privacy, and more specifically online privacy, seems like a zero-sum game: we’re trading privacy for convenience or better information. Is it now too late to combine the better of both worlds? Are we sufficiently aware of the consequences of our online actions? *Knowledge is power*: did we let some entities become too powerful?

8 REFERENCES

[1] Labotitz, C., “Google Sets New Internet Record”, Deepfield Labs, available on http://www.deepfield.net/blog/, last accessed 2013-11-09.

[2] Abraham, C., Nishihara, E., Akiyama, M., “Transforming healthcare with information technology in Japan: A review of policy, people, and progress”, International Journal of Medical Informatics 80, pp 157–170, 2011

[3] Hedge, A., James, T., Pavlovoc-Veselinovic, S., “Ergonomics concerns and the impact of healthcare information technology”, International Journal of Industrial Ergonomics 41, pp 345–351, 2011.

[4] Yin Ling Fung, M., Paynter, J., “The Impact of Information Technology in Healthcare Privacy, Ethical, Legal and Social Issues”, in Medical Informatics, pp. 186–227, 2008.

[5] Morosov E., “The Net Delusion The Dark Side of Internet Freedom”, Published in the United States by PublicAffairs™, 2011.

[6] Westin, A., “Privacy and Freedom”, New York, NY: Atheneum. 1967.

[7] Hage, H., Aïmeur, E., “Preserving Learners' Privacy”, in Advances in Intelligent Tutoring Systems by Riichiro Mizoguchi, Jacqueline Bourdeau and Roger Nkambou., Springer Verlag, pp. 465–482, 2010.

[8] Dropbox,“Privacy Policy – Security”, Available at https://www.dropbox.com/privacy#security. Last accessed 2013-11-08.

[9] Schneier, B., “Schneier on Security”, Wiley, 2009.

[10] Experian Marketing Services, “Experian Marketing Services Reveals 27 Percent of Time Spent Online is on Social Networking”, available at http://press.experian.com/United-States/Press-Release/experian-marketing-services-reveals-27-percent-of-time-spent-online-is-on-social-networking.aspx?WT.srch=PR_EMS_OnlineTime_041613_gpo , Last accessed 2013-11-09.

[11] McDonald, A. and Cranor, L., “The Cost of Reading Privacy Policies”, “I/S: A Journal of Law and Policy for the Information Society”, available on http://moritzlaw.osu.edu/students/groups/is/files/2012/02/Cranor_Formatted_Final.pdf, Last accessed 2013-11-09.

[12] Privacy Rights, “Online Information Brokers and Your Privacy”, PrivacyRights.org. Available at http://www.privacyrights.org/ar/infobrokers.htm . Last accessed 2013-03-05.

[13] Aïmeur, E., Schönfeld, D., “The ultimate invasion of privacy: identity theft”, Ninth Annual Conference on Privacy, Security and Trust, (PST-11) Montreal, pp. 24–31, July 2011.

[14] GeoNetwork, “The portal to spatial data and information”, Available at http://www.fao.org/geonetwork/srv/en/main.home. Last accessed 2013-11-08.

[15] BackCheck, “Your Social Media Footprint: Your Future Background Check Results”, Published 2013-06-03, available at http://www.backgroundcheckblog.com/your-

[8] https://fiber.google.com/about/

social-media-footprint-your-future-background-check-results/. Last accessed 2013-11-08.

[16] Enriquez, J., “TED Talks: Your online life, permanent as a tattoo”, Available at http://www.youtube.com/watch?v=Fu1C-oBdsMM&feature=youtu.be. Last accessed 2013-11-08.

[17] NetSplit, “IRC Networks – Top 100”, Available at http://irc.netsplit.de/networks/top100.php. Last accessed 2013-11-08.

[18] ComputerWorld, “New Snowden revelation shows Skype may be privacy’s biggest enemy”, Published on 2013-07-12, available at http://blogs.computerworld.com/privacy/22477/new-snowden-revelation-shows-skype-may-be-privacys-biggest-enemy. Last accessed 2013-11-08

[19] SiliconANGLE, “Skype Privacy? Doesn’t Exist, Sorry. Microsoft Can Read EVERYTHING”, Published on 2013-05-21, available at http://siliconangle.com/blog/2013/05/21/skype-privacy-doesnt-exist-sorry-microsoft-can-read-everything/. Last accessed 2013-11-08.

[20] Bloomberg Business Week Technology, “Cracking China’s Skype Surveillance Software”, Published on 2013-03-08, available at http://www.businessweek.com/articles/2013-03-08/skypes-been-hijacked-in-china-and-microsoft-is-o-dot-k-dot-with-it#p2. Last accessed 2013-11-08.

[21] Solove. D.J., “Privacy Self-Management and the Consent Dilemma”, to be published in 113 Harvard Law Review 2013.

[22] Solove, D.J., “A Taxonomy of Privacy”, GWU Law School Public Law Research Paper No. 129, 2006.

[23] Yun, S. and Pearson, S., “Privacy Enhancing Technologies: A Review”, HP Laboratories 2011-113 2011.

[24] Goldschlag, D. M. Reed, M. G. and Syverson, Paul F., “Hiding routing information”, in Information Hiding, pages 137–150. Springer-Verlag, 1996.`

[25] Reiter, M.K. and Rubin, A.D., “Crowds: anonymity for web transactions”, ACM Trans. Inf. Syst. Secur., 1(1):66–92, 1998.

[26] Dingledine, R., Mathewson, N. and Syverson, P., “Tor: The second-generation onion router”, In Proceedings of the 13th USENIX Security Symposium, pages 303–320, San Diego, CA, USA, August 2004.

[27] Levine, B. N. and Shields, C., “Hordes: A multicast based protocol for anonymity”, J. Comput. Secur., 10(3):213–240, 2002.

[28] Abadi, M., “Private authentication”, In Privacy Enhancing Technologies, volume 2482 of LNCS, pages 27–40, 2003.

[29] Cameron, K. and Jones, M. B., “Design rationale behind the identity metasystem architecture”, In ISSE/SECURE 2007 Securing Electronic Business Processes, 2007.

[30] Chaum, D., “Showing credentials without identification. signatures transferred between unconditionally unlinkable pseudonyms”, In EUROCRYPT ’85, pages 241–244, Springer-Verlag New York, NY, USA, 1986.

[31] Brands, S., “A technical overview of digital credentials”, International Journal on Information Security, 2002.

[32] Herzberg, A. and Mass, Y., “Relying party credentials framework”, Electronic Commerce Research, 4(1-2):23–39, 2004.

[33] Chaum, D., “Achieving electronic privacy”, Scientific American, pages 91–101, 1992.

[34] Agrawal, R. and Srikant, R., “Privacy-preserving data mining”, In Proc. of the ACM SIGMOD, pages 439–450. ACM Press, May 2000.

[35] Evfimevski, A., Gehrke, J. and Srikant, R., “Limiting privacy breaches in privacy preserving data mining”, In Proceedings of the ACM SIGMOD/PODS Conference, pages 211–222, San Diego, CA, June 2003.

[36] Chen, K. and Liu, L., “Privacy preserving data classification with rotation perturbation”, In ICDM ’05, pages 589–592, Houston, TX, November 2005.

[37] Kargupta, H., Liu, K. and Ryan, J., “Privacy sensitive distributed data mining from multi-party data”, In Proceedings of the first NSF/NIJ Symposium on Intelli- gence and Security Informatics, Lecture Notes in Computer Science, pages 336–342, Tucson, AZ, June 2003. Springer Berlin/Heidelberg.

[38] Machanavajjhala, A., Gehrke, J., Kifer, D. and Venkitasubramaniam, M., “l diversity: Privacy beyond k anonymity”, In 22nd IEEE International Conference on Data Engineering, Washington, pp. 26–27, 2006.

[39] Giannotti F., Atzori, M., Bonchi, F. and Pedreschi, D., “Blocking anonymity threats raised by frequent itemset mining”, In ICDM ’05, November 2005.

[40] Sweeney, L., “*k*-anonymity: A model for protecting privacy”, Int. J. Uncertain. Fuzziness Knowl.-Based Syst., 10(5):557–570, 2002.

[41] Pinkas, B., “Cryptographic techniques for privacy preserving data mining”, SIGKDD Explorations, 4(2):12–19, 2002.

[42] Gambs, S., Kégl, B. and Aïmeur, E., “Privacy-Preserving Boosting”, Data Mining and Knowledge Discovery, Vol. 14, pp. 131–170, 2007.

[43] Du, W., Han, Y. S. and Chen, S., “Privacy-preserving multivariate statistical analysis: Linear regression and classification”, In SDM ’04, Lake Buena Vista, FL, April 2004.

[44] Yang Wang, W., Komanduri, S., Leon, P. G., Norcie, G., Acquisti, A. Cranor, L.F., “I regretted the minute I pressed share”: A Qualitative Study of Regrets on Facebook” Symposium on Usable Privacy and Security (SOUPS) 2011, July 20–22, Pittsburgh, PA USA, 2011.

[45] Jacquemont, S., Jacquenet, F. and Sebban, M., “Sequence mining without sequences: A new way for privacy preserving”, In ICTAI ’06, pages 347–354. IEEE Computer Society, 2006.

[46] Matwin, S., Felty, M., Hernadvolgyi, I., T. and Capretta, V., “Privacy in data mining using formal methods”, In TLCA, pages 278–292, 2005.

[47] Cuzzocrea, A., Russo, V. and Saccà, D., “A robust sampling-based framework for privacy preserving olap”, In DaWaK ’08, pages 97–114, Berlin, Heidelberg, Springer-Verlag, 2008.

[48] Fung, B. C. M., Wang, Ke. and Yu, P. S., “Anonymizing classification data for privacy preservation”, TKDE, 19(5):711–725, May 2007.

[49] Zhelev, E. and Getoor, L., “Preserving the privacy of sensitive relationships in graph data”, In Proceedings of the First International Workshop on Privacy, Security, and Trust in KDD, pages 153–171, August 2007.

[50] Domingo-Ferrer, J., “A three-dimensional conceptual framework for database privacy”, In Secure Data Management, pages 193–202, 2007.

[51] Acquisti, A. and Gross, R., “Predicting social security numbers from public data”, in Proceedings of the National Academy of Science, vol. 106, no. 27, pp. 10975–10980, 2009.

[52] Golle, P. and Partridge, K., “On the anonymity of home/work – location pairs”, in Proceedings of the 7th International Conference on Pervasive Computing, Berlin, Palo Alto, pp. 390–397, 2009.

[53] Narayanan, A. and Shmatikov, V., “De-anonymizing social networks”, in 30th IEEE Symposium on Security and Privacy, Austin, pp. 173–187, 2009.

[54] Dalenius, T., “Finding a needle in a haystack—or identifying anonymous census record”, Journal of Official Statistics, vol. 2, no. 3, pp. 329–336, 1986.

[55] Loeb, “FTC Studies Online Targeted Advertasing; FTC Affiliate Marketing Rule”, loeb.com. Available at http://www.loeb.com/ftctargetedadvertisingaffiliatemarketing/. Last accessed 2013-03-10.

[56] Weinberger, D., “Everything Is Miscellaneous: The Power of the New Digital Disorder”, Henry Holt and Company, Business & Economics, 2008.

[57] Washington Post, “U.S., British intelligence mining data from nine U.S. Internet companies in broad secret program”, Available on http://www.washingtonpost.com/investigations/us-intelligence-mining-data-from-nine-us-internet-companies-in-broad-secret-program/2013/06/06/3a0c0da8-cebf-11e2-8845-d970ccb04497_story.html, last accessed 2013-11-09.

[58] Cloud Security Alliance, “CSA CloudBytes Town Hall: NSA/PRISM Lessons Learned”, Available on https://cloudsecurityalliance.org/media/news/cloudbytes-nsa-prism-lessons-learned/, last accessed 2013-11-09.

[59] Nevkontakte, “Google(tm) cache browser”, Available at http://cache.nevkontakte.com/. Last accessed 2013-11-08.

[60] Wikipedia, “The Slashdot Effect”, Available at http://en.wikipedia.org/wiki/Slashdot_effect. Last accessed 2013-11-08.

[61] Google, “Webmaster Tools – Request removal of a cached page”, Available at https://support.google.com/webmasters/answer/1663691?hl=en. Last accessed 2013-11-08.

[62] The Internet Archive, “The wayback machine”, Available at http://archive.org/web. Last accessed 2013-11-08.

[63] The Internet Archive, “Removing Documents From the Wayback Machine”, Available at http://archive.org/about/exclude.php. Last accessed 2013-11-08.

[64] The Independent, “EU court rules in Google’s favour: ‘right to be forgotten’ vetoed”, Published on 2013-06-25, available at http://www.independent.co.uk/news/world/europe/eu-court-rules-in-googles-favour-right-to-be-

forgotten-vetoed-8672512.html. Last accessed 2013-11-08.

[65] Brubaker, J., “Death and the social network: the persistence of digital identity”, Available at http://www.jedbrubaker.com/death-and-the-social-network-the-persistence-of-digital-identity/. Last accessed 2013-11-08.

[66] Prensky, M., “Digital Natives, Digital Immigrants”, On the Horizon (MCB University Press, Vol. 9 No. 5, 2001.