Privacy: whether you're aware of it or not, it does matter!

Bart Coppens, Olivier Zendra

There is growing awareness of the importance of privacy while, at the same time, we are sharing ever more private data with third parties. This creates an uneasy tension.

Privacy: whether you're aware of it or not, it does matter!

By BART COPPENS and OLIVIER ZENDRA

While privacy used to be a concern of only a limited number of people, in recent years awareness of it has been growing. This has been for a number of reasons including the enactment of the GDPR, the growing impact of data leaks, data logging by governments and companies, and even the recent discussions about COVID-19 contact tracing. At the same time, most of us are knowingly or unknowingly sending more and more private data to the cloud, which increases the risk of it being leaked or abused in some way.

In order to try and reconcile these two opposing directions, consumers and companies alike should increase their usage of privacy-enhancing technologies, and businesses should integrate privacy by design into their development.

Key insights

- Ever more data is being sent to and collected by governments and private companies alike.
- The scope and volume of the data being collected and analyzed is often not clear to consumers, who are even sometimes completely unaware.
- However, due to the GDPR and COVID-19, there is an increasing public awareness of privacy, not only of the fact that personal data is being collected, but also that it can be leaked, either on purpose or inadvertently.
- Technical solutions exist to improve privacy. The EU has a role to play.

Key recommendations

- The EU should promote research into technologies that enhance individuals' privacy and reduce the risks and impact of leaks of private data.
- The EU should encourage or even require companies to actively adhere to the principles of privacy by design.
- The EU should stand by its principles of privacy for its citizens, and not allow backdoors being put into applications.
- Existing solutions that limit leakage of personal information should be promoted by the EU.

We live in an era in which *almost everything we do is transmitted to servers beyond our control*. To give just a few examples, our private documents are stored in the cloud, while in some countries internet providers are legally obliged to keep track of which websites we visit [14]. Mobile service providers keep track of where our mobile phones make contact to their base stations and thus keep track of where we are; when we drive, our vehicle licence plates are captured by more and more automatic number-plate recognition (ANPR) cameras which are placed for various purposes by governments and municipalities [15,16,17]. The list goes on. People even freely put microphone-based listening devices such as Amazon Echo [11], Apple Siri devices [12], Google Nest [13], etc. in their homes for purposes of convenience and comfort.

Sometimes this sharing of information is quite intentional. As a matter of fact, sharing of information online has dramatically increased as a result of the sharp rise in home working caused by the COVID-19 pandemic lockdowns that also boosted e-commerce. When people do share a document online with others, they fully expect this information to be shared only with those specific individuals. However, the fact that this document is stored on servers – which can be hacked and can leak their documents – is something that most people forget. In addition, *most of the time people do not even realize the extent to which their private actions are tracked or shared with others*. People are surprised to find not only that their listening devices send out snippets of their private conversations to the companies that made their device, such as Amazon, but also that these private snippets are sent out to subcontractors who listen to them in order to increase the accuracy of the voice recognition engines that power these devices [18]. There is thus a real issue surrounding privacy and awareness of privacy issues.

It is thus clear that privacy is an important topic that directly affects the lives of many people. In the remainder of this article, we first discuss in more detail the kinds of **personal and private information** that nowadays are being generated, collected, and potentially leaked. We then describe

Image: ID 93286496 | © Valerii Brozhinskii | Dreamstime.com

some of the **technical directions** that can lead us to **protect** our **data and privacy** better.

Personal and private information

As a society, we are generating and storing ever-increasing amounts of (private) data. This includes the (confidential) data of companies. Almost all of this data, regardless of its source or whether or not it was sent intentionally by a user, is sent to and stored in cloud-based servers. This basically boils down to a consolidation of the software and hardware stacks of different users in the cloud. Because the infrastructure is shared between many different users, and is not located locally with these users, these cloud-based systems are much more vulnerable than locally run systems if they were to be unconnected from the network. Because both private customers and businesses need their cloud-based data to be secure from third-party snooping, leaking and interference, these systems need to be protected against many different kinds of attack.

Thanks to *regulations such as the General Data Protection Regulation (GDPR)*, European Union citizens should be better protected against at least some forms of unwanted processing of private data, and they should now at least be informed when such data is leaked or mishandled. This represents huge progress in terms of the public being informed and aware of data protection matters and should be hailed as a very positive step. Still, this does not mean that data leaks have magically gone away. For example, Figure 1 shows the number of data breaches involving US healthcare data, where in each case at least 500 records were leaked. The trend is unfortunately in the direction of more data breaches, not fewer.

Furthermore, even though we as European users of data platforms should now be *informed about the fact that data is collected, most people are unclear about the scope of the increasing amount of data that is being collected, processed, and stored*. The

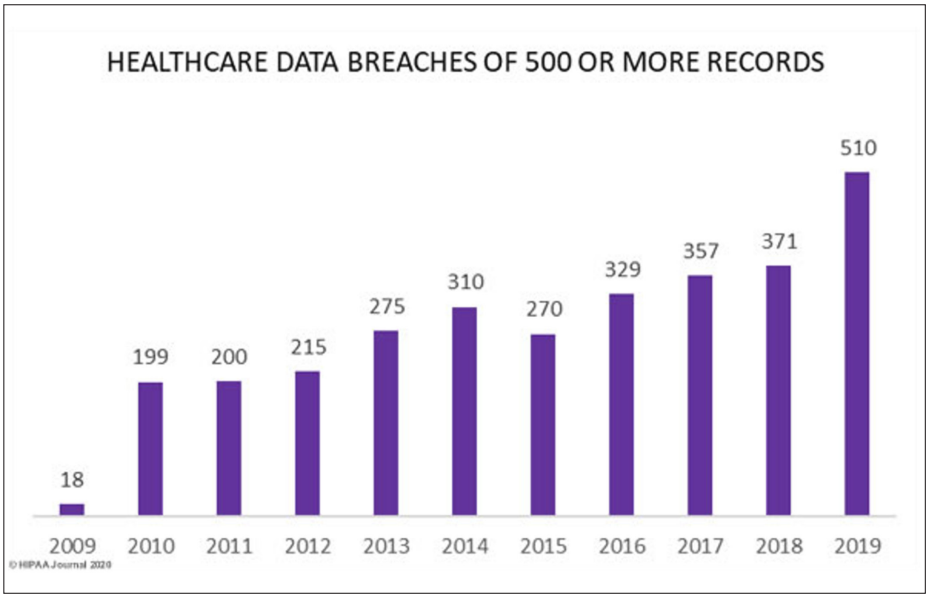

Figure 1: Number of breaches of 500 or more medical data records, as reported to the US Office for Civil Rights, Department of Health and Human Services (HHS) in the United States since October 2009 [1].

trend towards a service-based cloud economy only serves to exacerbate the scope to which this data is shared, and the risks to which this data is exposed.

While leaks of all kinds of data loom ever larger in the background of today's society, it typically remains a phenomenon that, in the mind of most people, is a concern only for others, rather than something that they feel will ever affect them (until it actually does, of course). Still, when (the threat of) a leak of private information does indeed seem imminent, sometimes it does indeed lead to more widespread debate. This happened only recently in the context of the COVID-19 pandemic. While contact tracing is something that had already happened for some other (but contained) infectious diseases such as tuberculosis, it had only been done manually and on a small scale by having contact tracing staff interview people directly. However, now that most people have a smartphone, technology has advanced enough to provoke proposals for contact tracing to be done by electronically tracking people's interactions with one another. In many countries, this sparked intense discussions about which design criteria such an app should adhere to: should it be based on location data and/or on Bluetooth-based proximity data, should it be totally anonymous or not, should the data be stored in a centralized or decentralized fashion, should it be mandatory or not, etc.

Despite these lengthy discussions on the very specific topic of contact tracing in the context of COVID-19, many people are still unaware of the extent to which data similar to this is already being kept track of. First and most obvious is the data that is already being kept that is related to government activities (ANPR cameras, cell phone data, ...). However, because of the privatization of the many functions of government, much of this data is already being kept by private companies.

Worse still, *even outside of such government-related activities, private companies are keeping track of increasingly more detailed information about people who are not even necessarily their users*. People's online activities are tracked by advertising companies such as Google and Facebook, so that they can then sell ever-more targeted ads [22,23]. They use data analysis techniques to create a profile of your interests, and even very private information such as your sexual orientation [28]. The users of such platforms are not always explicitly consenting to providing such information about their interests; user interests can be inferred implicitly using machine learning techniques which base themselves on the online behaviour of the tracked individuals in question. Some advertising companies go to quite some lengths to circumvent some anti-tracking measures that are implemented by browsers [24]. Some totally unscrupulous companies even scrape the internet for people's pictures from different social media channels and other websites, in order to build and train very accurate facial recognition of these people without their consent [25]. These facial recognition engines can then be sold to governments and commercial organizations across the world [26,27]. This tracking of data has further ramifications with regards to sovereignty. US or Chinese companies keep track of the data of EU citizens, harvest their pictures and process this data abroad, outside EU laws and regulations; this can create problems that are hard to address and solve.

Unfortunately, and quite surprisingly given the public debates over the COVID-19 tracking, no such sizeable debate is happening over these other kinds of even more widespread tracking and collecting of data. Still, we are hopeful that people will become increasingly aware of the fact that they do not want this kind of private and personal information to be used indiscriminately by parties beyond their control.

Technical means to protect our data and privacy

Now, how best to protect your data? The easiest solution would of course be to not share this data at all: unshared data truly is private data. However, the extent to which data is already being shared as soon as we try to interact with today's heavily digital society makes this infeasible except for people who are willing to become (partial) digital and social recluses. So, a middle ground needs to be found. This can be done at least in part by promoting and choosing technologies that enhance your privacy, rather than ignore it, or even worse, try actively to circumvent it.

In order to do so, we need to *design systems from the ground up with privacy in mind*. How a system handles private data and how it deals with privacy, should be design requirements from the start. In 2010, the International Conference of Data Protection and Privacy Commissioners published a resolution encouraging the recognition of the fact that privacy by design is an essential component of privacy protection, as well as the adoption of a set of foundational principles of privacy by design [29]. One of these principles is that privacy should be *embedded* in the design, and it should be an essential core of the functionality of the system [30]. These guiding principles were as true and necessary then as they are now, and their importance has only grown.

For example, rather than sharing documents and messages with people where the shared data is stored in an unencrypted form on cloud servers, we can try to use solutions with *true end-to-end encryption*. For example, one could use Signal [10] or EU-based Olvid [9] for sending messages to one another, rather than, for example, Facebook Messenger, as the former encrypt the data such a way that intermediate servers cannot decrypt the messages. In the latter case, Facebook has access to the original plain-text messages. When true end-to-end communications are not possible, or when people risk being tracked, it would still be beneficial to at least choose a *technology or solution which explicitly focuses on the privacy of its users*, like for example the EU-based Qwant search engine [31], or the Brave browser [32], that put an emphasis on protecting their users' privacy. The EU should encourage more initiatives and further developments and investments into such privacy-aware technologies and companies, to help protect its citizens and its sovereignty over data.

Of course, a good and easy way to have fewer problems with private data potentially being compromised is to not send it over the network at all. One way in which this

Credit: | ID 117352101 | ©Pop Nukoonrat | Dreamstime.com

can be solved is by having most or even all computations that would normally happen in the cloud, now happen locally, with the *processing being done at the edge*. This also has implications for industrial applications in the context of IoT and CPS: the more data is being processed in those devices themselves, rather than that the data has to be transmitted to cloud servers for processing, the less private data can be abused or leaked. A fog or federation of local devices sharing part of the global information in an encrypted way could solve the problem of accessing larger computing or storage resources in a more local manner.

If data does need to be transmitted or computed remotely, it is important to do this in a secure fashion that preserves as much security and privacy as possible. Most companies already try to *protect most sensitive data at rest and in transit with encryption*, for example with the Advanced Encryption Standard (AES) and Transport Layer Security (TLS). However, this data still needs to be processed, for which the data is currently still decrypted (and thus unprotected) on the systems that process it. Furthermore, if this data processing involves the data being searchable or queryable in a database, many systems will still store this data in an unencrypted form. One way to mitigate this problem is to *do the data processing on encrypted data*, in such a way that the personally identifiable information (PII) is not known to the system performing the actual processing. Examples of such techniques are (fully) homomorphic encryption (FHE), which still requires research to decrease its computing resource requirements, and secure multi-party computation. There are many fields in which homomorphic encryption would significantly increase the privacy of data in the presence of cloud-based data processing. In the medical sector, users would be able to upload their ECG data and have a cloud provider monitor their health without actually sharing their data with that cloud provider [2]. Similarly, we would be able to have our genome analyzed by third parties without information being passed on about which genetic diseases we have or other PII such as gender, race, etc [3]. Modifying different cloud-based machine learning tasks to protect PII would also significantly reduce the risks associated with outsourcing the relevant data. For example, face verification or face recognition would no longer expose photographs of people [4], and performing optical character recognition would no longer leak the text being processed [5].

Furthermore, if the recognized text is from licence plates that need to be queried in a database of stolen and wanted vehicles, for example, you can prevent the processing of all licence plates from leaking information about non-stolen cars [6]. The EU should invest in more technologies such as these, so that if PII data does need to be processed, the amount of data that can be intentionally or inadvertently leaked is minimized as much as possible.

Given the urgency for today's business landscape of the need to achieve more robust data privacy systems, we predict and advocate for an increase in the design and use of such homomorphic encryption and related techniques. Some start-ups already provide very specific applications of these techniques [7]. One limiting factor in applying FHE right now is its overhead. Both the time needed to process the data and the size of the messages that need to be exchanged with the cloud provider currently increase dramatically when FHE is applied. At present, this means that many of those techniques are unfortunately not yet usable in practice. In the meantime, some specific cases might not need to send the PII itself to third parties. Another issue to take into account when protecting data by encrypting it is how resistant the encryption scheme is to the changing landscape of attackers' capabilities. One clear but constant change is the increase in the processing speed of computers. As one of the most obvious goals of an attacker is to recover the information, the question is how long information can remain private, and how this time decreases with an increase in processing speed, and by how much we then increase the strength of the encryption (for example, by increasing the key size) to compensate for this. For traditional computers, it is quite clear how these scaling laws work, and increases in computing power do not immediately threaten the security of data encrypted with traditional encryption schemes. However, when switching to the different computing paradigm of quantum computers, this is not necessarily the case, because certain algorithms are believed to run significantly faster on quantum computers than on traditional computers. With some algorithms, it is sufficient to choose larger key sizes to compensate for this. However, other algorithms can be completely broken with quantum computers. Such algorithms need to be replaced with algorithms that could withstand attacks from a quantum computer. This field is called post-quantum cryptography.

However, it is not sufficient to use state-of-the-art encryption algorithms to protect PII. Software that is not secure can obviously leak all kinds of confidential and private information to attackers, even if under normal circumstances this data is stored and transmitted securely. Some *security-related Instruction Set Architecture extensions* have explicit implications for improving privacy. For example, one of the goals of Intel's Software Guard Extensions (SGX) is to protect the execution of certain code fragments from attackers that have control over the rest of the system, including the operating system itself. This can then be used to protect sensitive and private information even when the entire system is being attacked. However, the many recent attacks on SGX show that even this technology is clearly not yet mature enough to withstand such attacks in practice [19,20,21]. It may even be that the SGX model of allowing execution of code on private data, and general-purpose code execution by untrusted users, might not be feasible.

In this context, it is important to stress the importance of the entire system being *secure and not weakened by backdoors*. Some countries have argued for the presence of such backdoors in operating systems, telecommunications systems, and secure/encrypted communication platforms, such that only "they" can (lawfully) gain access to systems and decode encrypted information. These backdoors reduce the security of the entire system, since there is no guarantee that the law enforcing agents of your own country will be the only ones with access to these backdoors: other countries and criminals might be able to use them too. For example, a pseudo-random number generator containing a weakness in it which had allegedly been introduced by the NSA, eventually found its way into firewalls, where it was exploited by unknown parties [34]. Some people even claim that Intel's closed-source management engine on chips does not only support good-intentioned remote management features, but could also be used by other (malicious) parties to remotely gain access to machines [35].

These backdoors also reduce the overall trust people have in computers and telecommunication systems, thus undermining all the efforts of the EU to increase the privacy and security of its citizens. Not only that, such measures will also affect the confidential data of companies, which would then also become vulnerable to being leaked through these backdoors as well. Thus, to protect both the privacy of its citizens, and the confidential data of its companies, the EU should not give in to calls to action to even consider legalizing such backdoors.

However, while an insecure system can lead to information leaks, the converse is not necessarily true. A secure system cannot distinguish between purposeful leaks of information (for example, a user who wants to print his/her own bank statements), versus inadvertent leaks of information (for example, these bank statements being stored unencrypted on disk). One possible solution here is language-based information-flow security that allows programmers to explicitly define which flows of information are allowed, and to define properties on these flows [8].

A final source of leaking private information are the users themselves: often they are not aware of the actual private information that can be extracted from the data being shared: posting pictures of somebody in a bar or in a nightclub might be interpreted by an insurance company as somebody being a health risk because they drink alcohol. There are artificial intelligence-based systems that can analyse such public content and can warn users of this "side channel" information [33].

Conclusion

Public awareness of privacy issues is slowly increasing thanks to initiatives such as the enactment of the GDPR and to the issue of contact tracing for COVID-19. These will hopefully be a trigger for people

to think more about where and how their private data is being collected, stored, and used, which in many cases is anyplace, anytime, by most of the helper tools and applications (smartphones). This will hopefully lead people to try more actively to protect their own privacy. The EU has a role to play in terms of regulation and promoting and financing privacy and sovereignty preserving EU-based solutions.

References

[1] Healthcare Data Breach Statistics, HIPAA Journal, online, accessed December 3, 2020. https://www.hipaajournal.com/healthcare-data-breach-statistics/

[2] Kocabas, Ovunc, et al. "Assessment of cloud-based health monitoring using homomorphic encryption." 2013 IEEE 31st International Conference on Computer Design (ICCD). IEEE, 2013

[3] Miran Kim, Kristin Lauter, "Private genome analysis through homomorphic encryption", BMC Med Inform Decis Mak. 2015; 15(Suppl 5): S3.

[4] J.R. Troncoso-Pastoriza, D. González-Jiménez, F. Pérez-González, 2013. Fully private non-interactive face verification". IEEE Transactions on Information Forensics and Security, 8(7), pp.1101-1114

[5] Nathan Dowlin, Ran Gilad-Bachrach, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. 2016. "CryptoNets: applying neural networks to encrypted data with high throughput and accuracy. In Proceedings of the 33rd International Conference on International Conference on Machine Learning – Volume 48 (ICML'16), Maria Florina Balcan and Kilian Q. Weinberger (Eds.), Vol. 48. JMLR.org 201-210.

[6] Sunil, Archana Bindu, Zekeriya Erkin, and Thijs Veugen. "Secure matching of Dutch car license plates." Signal Processing Conference (EUSIPCO), 2016 24th European. IEEE, 2016

[7] https://www.privatebiometrics.com/, accessed December 4, 2020

[8] A. Sabelfeld and A. C. Myers, "Language-based information-flow security", in IEEE Journal on Selected Areas in Communications, vol. 21, no. 1, pp. 5-19, Jan. 2003.

[9] Olvid. https://olvid.io/technology/en/

[10] Signal. https://signal.org/docs/

[11] Amazon Echo. https://en.wikipedia.org/wiki/Amazon_Echo

[12] Apple Siri. https://en.wikipedia.org/wiki/Siri

[13] Google Nest. https://en.wikipedia.org/wiki/Google_Nest_(smart_speakers)

[14] Judgments in Case C-623/17, Privacy International, and in Joined Cases C-511/18, La Quadrature du Net and Others, C-512/18, French Data Network and Others, and C-520/18, Ordre des barreaux francophones et germanophone and Others. Press release 123/20, Court of Justice of the European Union, 6 October 2020

[15] Automatic Number Plate Recognition, Police.uk https://www.police.uk/information-and-advice/automatic-number-plate-recognition/ accessed December 2020

[16] Denmark: Targeted ANPR data retention turned into mass surveillance EDRi, September 6, 2017, https://edri.org/our-work/denmark-targeted-anpr-data-retention-turned-into-mass-surveillance/

[17] Automatic number-plate recognition - Usage, Wikipedia, https://en.wikipedia.org/wiki/Automatic_number-plate_recognition#Usage Accessed December 2020

[18] Apple contractors 'regularly hear confidential details' on Siri recordings, The Guardian, July 26, 2019.

[19] Foreshadow - Extracting the Keys to the Intel {SGX} Kingdom with Transient Out-of-Order Execution. USENIX Security Symposium 2018.

[20] Plundervolt: Software-based Fault Injection Attacks against Intel SGX. Murdock et al. IEEE Symposium on Security and Privacy 2020

[21] CrossTalk: Speculative Data Leaks Across Cores Are Real. Ragab et al. Accepted in the IEEE Symposium on Security and Privacy, 2021.

[22] Why targeted ads are the most brutal owns. Vox, September 25, 2018. https://www.vox.com/the-goods/2018/9/25/17887796/facebook-ad-targeted-algorithm

[23] Google's ad tracking is as creepy as Facebook's. Here's how to disable it. The Guardian, October 21, 2016. https://www.theguardian.com/technology/2016/oct/21/how-to-disable-google-ad-tracking-gmail-youtube-browser-history

[24] Ad Tech Surveillance on the Public Sector Web, Cookiebot Report, version July 14, 2020. https://www.cookiebot.com/media/1136/cookiebot-report-2019-ad-tech-surveillance-2.pdf

[25] Scraping the Web Is a Powerful Tool. Clearview AI Abused It. Wired, January 25, 2020. https://www.wired.com/story/clearview-ai-scraping-web/

[26] Clearview's Facial Recognition App Has Been Used By The Justice Department, ICE, Macy's, Walmart, And The NBA. Buzzfeed, February 27, 2020. https://www.buzzfeednews.com/article/ryanmac/clearview-ai-fbi-ice-global-law-enforcement

[27] Secret Users Of Clearview AI's Facial Recognition Dragnet Included A Former Trump Staffer, A Troll, And Conservative Think Tanks. Buzzfeed, March 11, 2020. https://www.buzzfeednews.com/article/ryanmac/clearview-ai-trump-investors-friend-facial-recognition

[28] Researchers Claim Facebook Ads Could Out LGBTQ+ Users. Out, August 30, 2019. https://www.out.com/tech/2019/8/30/researchers-claim-facebook-ads-could-out-lgbtq-users

[29] Resolution on Privacy by Design. 32nd International Conference of Data Protection and Privacy Commissioners. Jerusalem, Israel 27-29 October, 2010. https://edps.europa.eu/sites/edp/files/publication/10-10-27_jerusalem_resolutionon_privacybydesign_en.pdf

[30] Privacy by Design: The 7 Foundational Principles. Ann Cavoukian, Ph.D., Information & Privacy Commissioner of Ontario, Canada. https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf

[31] Qwant. Accessed December 2020. https://www.qwant.com/?l=en

[32] Brave. Accessed December 2020. https://brave.com/

[33] https://www.researchgate.net/publication/301221061_Personalized_Privacy-aware_Image_Classification

[34] Researchers Solve Juniper Backdoor Mystery; Signs Point to NSA, Wired, December 22, 2015. https://www.wired.com/2015/12/researchers-solve-the-juniper-mystery-and-they-say-its-partially-the-nsas-fault/

[35] Is the Intel Management Engine a backdoor? TechRepublic, July 1, 2016. https://www.techrepublic.com/article/is-the-intel-management-engine-a-backdoor/

Bart Coppens is Postdoctoral Researcher in the Electronics department of Ghent University, Ghent, Belgium.

Olivier Zendra is a Tenured Computer Science Researcher at Inria, Rennes, France.

This document is part of the HiPEAC Vision available at hipeac.net/vision.

This is release v.1, January 2021.

Cite as: B. Coppens and O. Zendra. Privacy: whether you're aware of it or not, it does matter! In M. Duranton et al., editors, HiPEAC Vision 2021, pages 88-93, Jan 2021.

DOI: 10.5281/zenodo.4719402

The HiPEAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871174.