The Role of Privacy-Preserving Technologies in the Age of Big Data

Daniel Bachlechner[1]
Competence Center Emerging Technologies, Fraunhofer ISI
Karlsruhe, Germany

Karolina La Fors
Center for Law and Digital Technologies, Leiden University
Leiden, The Netherlands

Alan M. Sears
Center for Law and Digital Technologies, Leiden University
Leiden, The Netherlands

ABSTRACT

The potential social and economic benefits of big data applications are highlighted by researchers and the media alike. However, they can also have negative implications, which are not limited to privacy issues. With alarming regularity, massive data breaches become public. Measures taken by both policy makers and business leaders do not seem to be effective. Privacy-preserving technologies have long been a hot topic in research, but they have not yet been widely integrated into big data solutions. To understand the mechanisms that drive or prevent the deployment of privacy-preserving technologies better, we investigated their effectiveness and the challenges they pose as well as their perception and use in the context of big data. The findings indicate that privacy-preserving technologies are quite mature, have different aims and need to be combined to be effective. The mechanisms that affect their deployment are manifold.

Keywords: Privacy, Data Protection, Technology, Big Data, Innovation

INTRODUCTION

Data-driven innovation is deeply transforming society and the economy. Although there are many potential social and economic benefits, big data applications can also have negative

[1] Corresponding author. daniel.bachlechner@isi.fraunhofer.de +49 721 6809 161

implications. Such implications are not limited to invasions of privacy but also include, for instance, discrimination and impairments regarding autonomy, solidarity or transparency. Privacy-preserving technologies (PPTs) tailored to the specifics of big data have received considerable attention in research over the last decade; however, the implications of this research on big data applications and related software solutions need further assessment. Most of the existing scientific literature dealing with PPTs focuses on technical details. Aspects relating to the integration of technologies in today's big data solutions are usually not considered at all. There are some publications, however, that discuss the advantages and disadvantages of PPTs as well as the related challenges and ways to overcome them with a focus on specific application contexts. Such publications contribute to the understanding of the role of the technologies in practice. Among the contexts addressed are, for instance, healthcare (Iyengar et al. 2018) and smart environments (Ziegeldorf et al. 2014). Through the lens of the Technology Innovation Systems (TIS) framework (Carlsson and Stankiewicz 1991), we investigated the diffusion of PPTs in the big data context. The TIS literature has generated valuable insights into factors relevant for the successful deployment of technologies. Because of the slow integration of PPTs into big data solutions, we paid particular attention to mechanisms that drive or prevent deployment and took both the structural components and functions of the innovation system into account. Studying both the effectiveness of PPTs and challenges related to them improved our understanding of the technological structures. The analysis of the perception and use of PPTs gave us insights into the institutional structures and the actors involved.

METHODOLOGY

The assessment of PPTs consisted of two parts: a technology-specific assessment focusing on the effectiveness of selected technologies and related challenges, and a more general assessment focusing on the perception and use of PPTs. We assessed technologies for anonymization and sanitization, encryption, multi-party computation (MPC), access control, policy enforcement, accountability, transparency, data provenance, access, portability, and user control. The assessment was based on semi-structured interviews and a review of related literature. We interviewed nine key informants, including researchers, company representatives and members of relevant organizations such as data protection authorities. The key informants were based in Europe, North America and the Middle East, and most of them had gained experience in multiple regions. The key informant interviews lasted between 30 and 45 minutes and were transcribed. Additionally, we conducted 16 short interviews, mostly with European researchers, which lasted up to 10 minutes. At the beginning of each interview, the technologies to be discussed (which are introduced in the next section) were briefly outlined. Afterwards, the interviewees were asked about the effectiveness of the technologies and challenges related to them. After that, we asked questions about the perception and use of the technologies. We conducted a content analysis of the transcripts following the methodology proposed by Krippendorff (2013). The findings of the analysis were validated by means of the short interviews and related to the literature.

TECHNOLOGIES

This section introduces the studied classes of technologies. Anonymization is performed by encrypting or removing personally identifiable information from datasets. Examples for privacy models that may be used include k-anonymity and differential privacy. Sanitization is done by encrypting or removing sensitive information from datasets. Anonymization is a type of sanitization. Encryption is the encoding of information so that only authorized parties can access it. Examples for cryptographic primitives that are relevant in the context of big data include attribute-based encryption, functional encryption and homomorphic encryption. MPC relies on the distribution of data and processing tasks over multiple parties. MPC is a field of cryptography that aims to allow securely computing the result of functions without revealing the input data. Access control describes the selective restriction of access to resources. Attribute-based access control refers to a set of approaches that support fine-grained access control policies based on attributes that are evaluated at run-time. Policy enforcement focuses on the enforcement of rules for the use and handling of resources. Data expiration policies, for instance, are already enforced by some big data solutions. Accountability requires the evaluation of compliance with policies and the provision of evidence. A cornerstone of accountability in the context of big data is the provision of automated and scalable control, as well as auditing processes that can evaluate the level of compliance. Transparency calls for the explication of information collection and processing. In the age of big data, transparency is achieved by multichannel and layered approaches as well as standardized icons. Data provenance relies on being able to attest the origin and authenticity of information. Access and portability facilitate the use and handling of data in different contexts. Having access to data means that people can view the data stored. Portability gives people the possibility to change service providers without losing their data. User control refers to the specification and enforcement of rules for data use and handling. Consent mechanisms are one means that enables far-reaching user control, others are privacy preferences and personal data stores.

EFFECTIVENESS AND CHALLENGES

In the first part of the assessment, we paid particular attention to the effectiveness of the technologies in protecting privacy, and challenges that arise when using the technologies.

Anonymization and Sanitization

Anonymization and sanitization technologies are very relevant in the big data context since data cannot be controlled anymore as soon as it has been released. It is important that measures are taken to reduce the risk that people are re-identified or that sensitive attributes are inferred. There is a substantial amount of literature on technologies for anonymization and sanitization. Technologies can be used to protect, anonymize or aggregate data in ways that are effective and efficient (Acquisti and College 2010). The key challenge is to determine the optimal balance between improved privacy protection and the usefulness of the data for decision-making. On the one hand, the data should be utilized for data mining and extracting value, and, on the other hand, the re-identification of the data should be at least very hard, if not impossible. Sometimes it may be that there is no satisfying trade-off: either some utility and very weak privacy, or some privacy and hardly any utility. Further challenges are, for instance, the uniqueness of certain characteristics or behaviors, and the fact that it is usually unknown what other information is available to a potential adversary. Technologies based on differential privacy are most promising. Sanitization is good only to prevent accidental disclosure, not to provide protection from a motivated adversary. A key drawback is that there are many examples where anonymization has failed. Anonymization is particularly difficult with respect to medical data. The difficulties are caused, for instance, by free text that includes names, indirect descriptions of things such as diseases or treatments, or dates that can easily be cross-related with other data sources. Therefore, purpose limitation has particular relevance in this context.[2] The mathematical background for anonymization is established and very stable. However, guidelines for the choice of parameters such as ϵ in differential privacy are missing (Karnouskos and Kerschbaum 2018).

Encryption and Multi-Party Computation

Encryption is generally strong and fundamental for the protection of data (D'Acquisto et al. 2015). As cloud storage is increasingly used, encryption is crucial to ensure the

[2] Purpose limitation is a principle that prevents an organization from using personal data for new purposes if they are 'incompatible' with the original purpose for collecting the data. GDPR, Art. 5(1).

confidentiality and integrity of sensitive information (Al Mamun et al. 2017). In the context of big data, it is necessary to go beyond the ‘encrypt all or nothing’ model. With respect to encryption, it is important to keep the trust model in mind. As long as fully trusted parties are exchanging encrypted data and related keys, everything is fine. If the parties do not fully trust each other, encryption does not provide any protection. An approach that is considered to maximize privacy and query expressiveness, at least theoretically, is fully-homomorphic encryption (FHE). A key challenge with respect to FHE is computation cost that makes it relatively slow compared to other methods. Therefore, further research must be performed. If zero quality loss is not the number one requirement, there are other more interesting approaches. Semi-homomorphic encryption, for instance, is mature enough to be integrated into products. Secure MPC is less secure than FHE, especially when many untrusted parties are involved but more efficient in certain types of implementations (Danezis et al. 2014). MPC has a long history, but practical implementation problems stand in the way of wide adoption. Research can help by addressing the problems, for instance, by providing algorithms that speed up common analytics methods. MPC can unlock new possibilities in the context of joint data processing in domains where legal and procedural hurdles are prevalent. While anonymization allows for the use of standard analytical tools, the value of the data may be decreased in relation to the scope of the anonymization process. MPC enables working with the data as it is, but it restricts the efficiency of the analysis and the range of tools that can be used.

Access Control and Policy Enforcement

Big data applications typically require fine-grained access control. One of the interviewees reported that the software and services company he works for has put a lot of engineering effort into developing an access control model that allows organizations to provision access in a highly granular and dynamic way. With respect to granularity, the company allows granting access to data all the way down to the sub-cell level for data in tabular format. With respect to dynamics, access is granted for each session based on the user's role as well as the needs of the specific task the user is performing. For instance, a case management system used to do policing work grants access not only based on the user's general status but also taking the severity of the crime into account. This allows restricting access to privacy-invasive datasets in case of minor crimes.

Some classes of technologies including those focusing on access control and policy enforcement are threatened by a single point of failure and limited transparency. Usually, there is one person that designs the access control system and one that specifies the policies. Opacity is often intensified by trade secrets or practices that are not completely open. In such cases, non-technical measures such as specific processes complementing or replacing technologies become relevant, albeit sometimes cumbersome. For data protection authorities, it is particularly hard to exercise their enforcement power if foreign actors are involved. Policy enforcement technologies play a key role in the context of big data as chains of responsibility become longer and more geographically dispersed. Enforcement frameworks must be flexible and able to support different data processing requirements (Inukollu et al. 2014). Automated policy enforcement mechanisms are important in the big data era as policies get lost easily when data is transferred between different systems. While access control technologies have a long history, technologies for policy enforcement are not yet mature.

Accountability, Transparency and Data Provenance

Accountability and transparency technologies are highly relevant in the big data context. For instance, if a classifier is trained using a large dataset, it may be necessary to ensure that the classifier does not discriminate against particular groups. Therefore, a measure for fairness and a way to ensure that the learning is fair are needed. Moreover, explanations of certain decisions may be needed. For example, if a classifier is used to decide whether somebody receives a loan or not, it would be good to have an explanation of the decision. In the big data context, it is extremely difficult to explain what algorithms do with data or to get a preview of what the outcome of providing data may be (Diakopoulos and Friedler 2016). Machine learning modules typically are 'black boxes'. If an Internet of Things (IoT) device performs in an undesired way, it is often almost impossible to find out why; the performance is not interpretable. There is a growing field of research that aims to explain why certain decisions are made. A key problem is that most IoT devices do not have a screen, which makes it difficult for them to visually explain what they are doing. Accountability and transparency technologies are not yet mature.

Big data poses challenges for data provenance (Wang et al. 2015). Problems are caused by the strong heterogeneity of the data. Additionally, the use of many analytics and storage solutions may result in prohibitively large amounts of provenance information. One of the interviewees pointed out that the company he works for offers products that allow not only integrating data from different sources but also preserving the sources of all components in a unified model. Each property that is related to an object can come from a different source. The interviewee stressed that the company puts a lot of development effort into notions of data provenance and underlined the tight connection between data provenance, accountability and transparency. The measures taken by the company with focus on data provenance allow users to investigate values that seem wrong. They cannot only check if, how and when errors were introduced but also implement a fix and, based on the provenance tree, rebuild the dataset with the corrected values. Provenance technologies are quite mature.

Access, Portability and User Control

Access and portability foster competition. However, they can create problems as data may be brought from one domain where there are safeguards to another domain that is riskier. Moreover, access and portability need trust between the involved parties. A user needs to be sure that whenever he or she brings data from one place to another, the data is processed according to his or her expectations. This cannot be done without a prior agreement between the parties. As service providers do not only face a higher risk of losing users but also see the opportunity to gain users, many will eventually accept the paradigm of access and portability. With respect to user control, the informational asymmetry between users and data collectors is considered a central challenge. The fact that people get control over their data does not resolve the asymmetry. The user cannot understand what his or her data can be used for and how certain database transactions will end up being inefficient or unfair. The fact that the user has greater control does not change this. Consequently, providing the user with more information to make the consent more informed does not really address the underlying challenge. People should be empowered, but to some extent, this also means that responsibility is pushed onto them. Therefore, big data solutions should be designed in ways that make it difficult for users to endanger themselves rather than just giving them controls and expecting them to know how to use them. Even for experts it is sometimes difficult to understand what the outcome of certain decisions will be, especially where complicated algorithms are used. Technologies for user control are currently neither mature nor user-friendly enough to be used by millions of people.

PERCEPTION AND USE

The second part of the assessment was more general and focused on the perception and use of the technologies. We paid particular attention to the integration of the technologies into today's big data solutions, the demand for big data solutions that include PPTs, related regional differences, technological boundaries, and the societal responsibility for privacy protection.

Technology Deployment

PPTs are integrated into today's big data solutions only to a very limited extent. The lack of privacy-preserving big data solutions is a big impediment for the data economy. On the one hand, incidents such as the one involving Cambridge Analytica and Facebook are likely to increase the reluctance of companies to share data. On the other hand, companies may refrain from attaching much importance to privacy because they fear that they could lose some of the benefits of using personal data (Altman et al. 2018). As protecting privacy is sometimes at odds with business objectives, developers do not integrate everything they can into their products. A change in culture would be needed. New legislation as well as incidents reported upon by the media, may help to bring about such a change.

Privacy by design requires privacy safeguards to be integrated into solutions. However, the principle has not yet arrived in practice. Companies often still reside in the early stages of privacy program development and have not yet set up the required design processes (Solove 2018). Another factor that might hamper the integration of the technologies into big data solutions is complexity. Developers often do not have the training needed. Particularly in regards to encryption technologies, there are strong research results, but there is a large gap when it comes to deployment. According to one of the interviewees, companies appear to use encryption technologies to protect their own datasets rather than to protect their users from themselves. Additionally, companies seem to prefer non-technical over technical measures as they interfere less with their need for flexibility regarding the use of data.

Demand for Privacy Protection

Currently, there seems to be little demand for big data solutions that include PPTs. Although people should generally be concerned about their privacy, they often appear to be indifferent unless something went wrong and the media has reported on the incident. Regulators could step in and make sure that organizations are transparent. According to one of the interviewees, organizations processing personal data must go beyond telling people what they do; they need to enable them to actually see what they do. The right to retract consent would be very artificial if people are not able to understand how their data is processed. People who do care, frequently do not know what to do. Technologies and concepts are often complex and counter-intuitive. Moreover, people are not used to the adversarial thinking required to understand threats. Younger people might have less concerns than people that are in their thirties or older. Something that demonstrates that there is at least some demand is that companies have begun using privacy protection as a selling point. Companies are distinguishing themselves from others by emphasizing that they do not monetize user data or that they made a particular effort to preserve the privacy of users. Demand is closely related to the maturity of the technologies. In any case, PPTs would have to be embedded into products rather than provided as add-ons. To prevail, privacy preservation has to lead to win-win situations. It is unlikely that people would pay extra for privacy preservation (Beresford et al. 2012). Policy makers could play an important role with respect to the demand for privacy protection. They could set priorities accordingly in education, take regulatory measures or put emphasis on privacy in public procurement.

Regional Differences

There are considerable regional differences with respect to the perception and use of PPTs in the big data context. We paid particular attention to the differences between Europe and

North America. Whereas Europe currently seems to be leading the way in terms of data protection legislation, most of the affected technological innovations do not come from there. In North America, which in recent history has been the birthplace of many technological innovations, companies tend to have fewer constraints in the early stages of their life cycle. One interviewee, a professor at a North American university, stated that the approach there is to see what happens and then to generalize and apply case-based legal decisions. The European historical context is generally more rule driven. While Europe may not have produced the same number of technological innovations as North America, its legislation often affects organizations far beyond Europe and thus has extraterritorial effect. Google, for instance, changed its privacy policy and its mechanisms for obtaining consent from its users because of the General Data Protection Regulation (GDPR). The company, according to one of the interviewees, had discovered that there was no way out. In the medium run, it is possible that the EU will become an exporter of norms that have the potential to lead to technological changes globally. However, it is also possible that certain cutting-edge technologies may not access, or have delayed access to the European market due to regulations. Apple, for instance, stated that it is modifying its products to comply with the GDPR, and the modification will be worldwide for everyone.

Technological Boundaries

A combination of technical and non-technical measures is essential. For example, organizations need to check the legal compliance of their own products and services before they are rolled out, and to deal with aspects such as consent. To ensure privacy, an effective regulatory framework and proper processes are essential complements to technical measures. Recent data breaches, however, clearly show the limitations of reactive approaches that typically prescribe measures to be taken when privacy is violated or when certain rules are broken. A growing interest in technologies that are proactive in the sense that they prevent incidents or rule violations in the first place can be seen. Moreover, there is a particular need for increased awareness, improved usability and education targeting, for instance, young people and users of social networks. Currently, most of the technologies can only be used correctly by experts. Making data a bit more difficult to re-identify usually requires it to be processed in such a way that it loses utility. It is a trade-off and technology cannot yet identify the 'sweet spot' entirely by itself. In general, technologies alone are not sufficient. At some level, non-technical measures will always be necessary to make sure a given technology functions as expected. Moreover, data protection officials are needed that are aware and able to assess the impact on privacy or risks regarding personal data that is collected and used by companies.

Societal Responsibility

With respect to the distribution of the responsibility for protecting privacy in the era of big data, there is no commonly agreed-upon position. However, eventually everyone should be working together and be responsible in line with his or her competency. The data controller and the data processor, as the strongest parties, should bear the largest responsibilities, especially if it is through their infrastructure that the data may be compromised. Data subjects should raise awareness and exercise their rights because this results in pressure on controllers and processors to respect user expectations and meet legal requirements. It is important that the responsibility is not pushed to the people who may not fully understand what they are doing. The responsibility placed on the user should be as small as possible. One of the interviewees emphasized that it would be good for people to have a way to anonymize or sanitize data before it leaves their sphere of control. Supervisory authorities and governments are important because they shape the framework conditions and can help enforce the rights of users. Developers of the big data solutions should be aware of the privacy risks and integrate some of the discussed technologies. It is then up to the users of the solutions to actually use them responsibly.

CONCLUSIONS

The assessment of PPTs led to the following overarching observations: The set of studied classes of technologies is quite comprehensive. Additional technologies were suggested by the interviewees, but most of them were at least related to the existing classes. The technologies contribute to privacy preservation in different ways and are closely linked with each other. In practice, the technologies need to be combined to be effective and there is no single most important class of technologies. The technologies pursue different aims. While some aim at overcoming the need for trust, others aim at increasing trust in other parties. A multidimensional measure for privacy preservation is needed that covers relevant factors in a balanced way. Apart from the degree and cost of protection, it is important to also take the societal value of data into account. Some interviewees reported a fundamental tension between the objectives of big data and privacy; they seem almost antagonistic.

There is wide agreement that PPTs are seldom integrated into today's big data solutions. There is very promising research, but there is a large gap when it comes to deployment. There is broad consensus with respect to the rather low demand for technologies that protect privacy. Regional differences, however, need to be taken into account. There is consensus that the technologies need to be complemented by non-technical measures. For modern organizations, awareness and education are aspects that are just as important as processes, legislation and policies. It is often stressed that people need to protect themselves because one cannot currently rely on someone else to do it for them. There is wide agreement, however, that the strongest party should have the largest responsibilities.

ACKNOWLEDGEMENT

The research leading to the presented results has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 731873.

REFERENCES

Acquisti, A., and College, H. 2010. "The Economics of Personal Data and the Economics of Privacy: 30 Years after the OECD Privacy Guidelines," *Background Paper* 3, OECD.

Al Mamun, A., Salah, K., Al-maadeed, S., and Sheltami, T. R. 2017. "BigCrypt for Big Data Encryption," in *Proceedings of the 4th International Conference on Software Defined Systems.* 8-11 May 2017, Valencia, Spain: IEEE, pp. 93–99.

Altman, M., Wood, A., O'Brien, D. R., and Gasser, U. 2018. "Practical Approaches to Big Data Privacy Over Time," *International Data Privacy Law* (8:1), pp. 29–51.

Beresford, A. R., Kübler, D., and Preibusch, S. 2012. "Unwillingness to Pay for Privacy: A Field Experiment," *Economics Letters* (117:1), pp. 25–27.

Carlsson, B., and Stankiewicz, R. 1991. "On the Nature, Function and Composition of Technological Systems," *Journal of Evolutionary Economics* (1:2), pp. 93–118.

D'Acquisto, G., Domingo-Ferrer, J., Kikiras, P., Torra, V., de Montjoye, Y.-A., and Bourka, A. 2015. "Privacy by Design in Big Data: An Overview of Privacy Enhancing Technologies in the Era of Big Data Analytics," ENISA.

Danezis, G., Domingo-Ferrer, J., Hansen, M., Hoepman, J.-H., Le Métayer, D., Tirtea, R., and Schiffner, S. 2014. "Privacy and Data Protection by Design: From Policy to Engineering," ENISA.

Diakopoulos, N., and Friedler, S. 2016. *How to Hold Algorithms Accountable*. https://www.technologyreview.com/s/602933/how-to-hold-algorithms-accountable/. Accessed 11 September 2018.

Inukollu, V. N., Arsi, S., and Rao Ravuri, S. 2014. "Security Issues Associated with Big Data in Cloud Computing," *International Journal of Network Security & Its Applications* (6:3), pp. 45–56.

Iyengar, A., Kundu, A., and Pallis, G. 2018. "Healthcare Informatics and Privacy," *IEEE Internet Computing* (22:2), pp. 29–31.

Karnouskos, S., and Kerschbaum, F. 2018. "Privacy and Integrity Considerations in Hyperconnected Autonomous Vehicles," *Proceedings of the IEEE* (106:1), pp. 160–170.

Krippendorff, K. 2013. *Content Analysis*: *An Introduction to its Methodology*, Los Angeles, London, New Delhi, Singapore: Sage.

Solove, D. 2018. *Strategic Privacy by Design: An Interview with Jason Cronk*. https://teachprivacy.com/strategic-privacy-by-design/. Accessed 12 September 2018.

Wang, J., Crawl, D., Purawat, S., Nguyen, M., and Altintas, I. 2015. "Big Data Provenance: Challenges, State of the Art and Opportunities," in *Proceedings of the 2015 IEEE International Conference on Big Data,* Santa Clara, CA, USA. 29 October-1 November 2015, IEEE, pp. 2509–2516.

Ziegeldorf, J. H., Morchon, O. G., and Wehrle, K. 2014. "Privacy in the Internet of Things: Threats and Challenges," *Security and Communication Networks* (7:12), pp. 2728–2742.