Software Engineering for Privacy in-the-Large

Pauline Anthonysamy*† and Awais Rashid*
*Security-Lancaster Research Centre
Lancaster University, UK
{p.anthonysamy, a.rashid}@lancaster.ac.uk
†Google, Zurich, Switzerland
{anthonysp}@google.com

***Abstract*—There will be an estimated 35 zettabytes (35×10^{21}) of digital records worldwide by the year 2020. This effectively amounts to privacy management on an ultra-large-scale. In this briefing, we discuss the privacy challenges posed by such an ultra-large-scale ecosystem - we term this "Privacy in the Large". We will contrast existing approaches to privacy management, reflect on their strengths and limitations in this regard and outline key software engineering research and practice challenges to be addressed in the future.**

I. Introduction to the Topic and Relevance to Software Engineering Community

User data and personal information is shared and exists in a complex ecosystem that is not just limited to large-scale online social networks but also includes a wide range of other social activities, e.g., social search; social shopping; location-based access; and usage of cloud services, etc. (collectively referred to as online social media). There will be an estimated 35 zettabytes (35×10^{21}) of digital records worldwide by the year 2020. This effectively amounts to privacy management on an ultra-large-scale. Software engineers are modelling and developing such systems on a regular basis – these systems are increasingly more and more open and handle large amounts of personal or other sensitive data. Hence, software engineers and the tools and techniques at their disposal must address the privacy in-the-large needs inherent to such systems.

This technical briefing provides a synthesis of the state-of-the-art techniques in software engineering and some related technologies that can aid software engineers in tackling privacy in-the-large. At the same time, the briefing also highlights the new frontier presented by these ultra-large-scale settings and key software engineering research and practice challenges to be addressed in the future. Researchers will benefit from a synthesis of the state-of-the-art in this area while practitioners will benefit from an overview of the existing set of tools and techniques at their disposal as well as their maturity with regards to handling these challenges. Both audiences will benefit from discussion of open problems and the challenges as well as the opportunities they present for next generation software engineering approaches.

II. Topics to be covered in the Briefing

Although software engineering research has produced impressive results over the past years, new major trends in information technology lead to an amplification of existing challenges as well as the emergence of novel challenges. The massive collection, processing and dissemination of information in hyper-connected settings have led to privacy concerns regarding potential individual and societal harms. To curb these concerns, significant effort has been put into the research and development of privacy enhancing software engineering tools and approaches. The technical briefing will discuss key software engineering challenges posed by privacy in-the-large and relevant approaches from four different perspectives:

1) Privacy from the Perspective of Compliance
2) Privacy from the Perspective of Access Control
3) Privacy from the Perspective of Verification
4) Privacy from the Perspective of Usability

Privacy from the Perspective of Compliance: Increasingly, regulations in Europe and the US are governing the use and disclosure of user information in both industry and government. If an organisation uses, processes or stores personal information, then it is obligated to comply with the laws and regulations prescribed for that specific domain. In Europe, for instance, the EU's Data Protection Directive requires organisations that handle user information, e.g., web service providers, to comply with seven principles – *Notice, Purpose, Consent, Security, Disclosure, Access, and Accountability* – that govern the protection of personal data. This presents a variety of challenges to the business practices of organisations that handle user data. This part of the briefing will present research efforts that have focused on ensuring software system compliance by deriving privacy requirements from regulatory documents. Examples of such approaches include those that focus on eliciting and analysing requirements that are necessary to develop systems, such as health-care, hotel management, etc., which are data protection legislation compliant [1], [2], [3]. Whereas, other approaches focus on expressing traceability relationships between various software entities such as legal documents, and source code [4], [5].

Privacy from the Perspective of Access Control: The access control perspective of privacy dictates solutions that prevent abuses of data that is being collected. There has been a wide range of tools and mechanisms that provide access control for individuals and service providers by defining detailed rules to govern access to personal information. These include privacy enhancing technologies such as anonymity and anonymisers [6], differential privacy [7] and privacy-preserving data

publishing [8]. This part of the briefing will discuss such approaches and their effectiveness or otherwise with regards to privacy in-the-large.

Privacy from the Perspective of Verification: In general terms, verification is a process of establishing the truth, accuracy, or validity of something. From the perspective of privacy, verification is primarily directed at the process of checking that a software system, namely one that handles personal data, meets certain specifications or rules and that it fulfils its intended purpose. This part of the briefing will cover various approaches to formally modelling and exploring system models to verify privacy properties. Some examples include model checkers and model-based approaches [9], [10], [11]. They innovate by applying formal methods for verification of security and privacy properties which enhances software reliability thereby, increasing the usability of systems that employ them.

Privacy from the Perspective of Usability: This perspective focuses on research in terms of usability and interaction – how privacy solutions are perceived in a user context. This part of the briefing will cover research at the boundary of software engineering and human-computer interaction (HCI) to improve comprehensibility and usability of privacy preserving mechanisms. Usability researchers focus on the evaluation and understanding of user behaviours, needs, and motivations through observation techniques, and analysis of usability problems of existing privacy solutions. This perspective covers a wide spectrum which includes user studies on privacy perceptions [12], [13], privacy breaches in social media [14], and improvement of user awareness [15].

Future Challenges and Opportunities: The briefing will be concluded with a synthesis of the above four perspectives leading to open challenges to be addressed by software engineering research and practice. Examples of such future challenges include: privacy policy formulation and enforcement in complex topologies that crosscut platform, system and organisational boundaries; real-time compliance monitoring; access control models for partially-trusted information eco-systems; and development of privacy-preserving software architectures for data mining and analytics.

III. Author Biographies

Pauline Anthonysamy is a Privacy Researcher at Google and a member of the Security Lancaster Research Centre at Lancaster University, UK. Her research interests include data-driven privacy research, development of software engineering methods and tools for privacy policy compliance and traceability to a system's runtime functionality and the modelling and analysis of privacy in online social networks. Anthonysamy received her PhD in Computer Science from Lancaster University, United Kingdom. Her article titled "*Social Networking Privacy: Understanding the Disconnect from Policy to Controls*" was selected to be featured by the IEEE Special Technical Community on Social Networks. She is also a Google Anita Borg Scholar 2012 and was selected as one of 200 young researchers to attend the Heidelberg Laureate Forum 2014. She is a member of the ACM. Note: This research was done by her prior to her joining Google.

Awais Rashid is a Professor of Software Engineering in the School of Computing and Communications and Director of the cross-disciplinary Security-Lancaster Research Centre at Lancaster University, UK. He has been an active researcher in software engineering since the late 90s, mainly working on requirements engineering, modularity and variability management approaches, and more recently security and privacy challenges for large-scale systems. He has served on program committees of various conferences such as RE, FSE, ECOOP, AOSD, MODELS and is serving on the programme committee of the (new) ICSE Software Engineering in Society track. He has also given tutorials at AOSD, RE, ICSE and various other conferences. He has also been engaged in delivering policy and practice briefings on security and privacy issues and was the lead author of the UK Case Study for the UN's Internet Governance Forum in 2009. He is a member of the IEEE and the IEEE Computer Society.

References

[1] T. Breaux and A. Antón, "Analyzing regulatory rules for privacy and security requirements," *IEEE Trans. Softw. Eng.*, vol. 34, pp. 5–20, January 2008.

[2] A. Massey, P. Otto, L. Hayward, and A. Antn, "Evaluating existing security and privacy requirements for legal compliance," *Requirements Engineering*, 2010.

[3] J. Young, "Commitment analysis to operationalize software requirements from privacy policies," *Requirements Engineering*, 2011.

[4] J. Cleland-Huang, A. Czauderna, M. Gibiec, and J. Emenecker, "A machine learning approach for tracing regulatory codes to product specific requirements," in *ICSE*, 2010.

[5] G. Antoniol, G. Canfora, G. Casazza, A. De Lucia, and E. Merlo, "Tracing object-oriented code into functional requirements," in *Program Comprehension, 2000. Proceedings. IWPC 2000. 8th International Workshop on*, 2000, pp. 79–86.

[6] A. Pfitzmann and M. Hansen, "Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management – a consolidated proposal for terminology," http://dud.inf.tu-dresden.de/Anon_Terminology.shtml, Feb. 2008, v0.31.

[7] C. Dwork, "Differential privacy," in *in ICALP*. Springer, 2006.

[8] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information," in *Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, ser. PODS '98. New York, NY, USA: ACM, 1998.

[9] E. M. Clarke, Jr., O. Grumberg, and D. A. Peled, *Model Checking*. Cambridge, MA, USA: MIT Press, 1999.

[10] K. Fisler, S. Krishnamurthi, L. A. Meyerovich, and M. C. Tschantz, "Verification and change-impact analysis of access-control policies," in *Proceedings of the 27th International Conference on Software Engineering*, ser. ICSE '05. New York, NY, USA: ACM, 2005, pp. 196–205.

[11] D. Basin, F. Klaedtke, and S. Müller, "Monitoring security policies with metric first-order temporal logic," in *Proceedings of the 15th ACM Symposium on Access Control Models and Technologies*, ser. SACMAT '10. New York, NY, USA: ACM, 2010, pp. 23–34.

[12] P. B. Brandtzaeg and M. Lüders, "Privacy 2.0: Personal and consumer protection in new media reality," Tech. Rep. SINTEF A12979, Nov'09.

[13] M. L. Johnson, S. Egelman, and S. M. Bellovin, "Facebook and privacy: it's complicated," in *SOUPS*, 2012, p. 9.

[14] S. Gurses, R. Rizk, and O. Gunther, "Privacy design in online social networks: Learning from privacy breaches and community feedback," in *ICIS 2008 Proceedings*. New York, USA: ACM, 2008.

[15] A. Acquisti and R. Gross, "Imagined communities: Awareness, information sharing, and privacy on the facebook," in *Privacy Enhancing Technologies*, ser. Lecture Notes in Computer Science, G. Danezis and P. Golle, Eds. Springer Berlin Heidelberg, 2006, vol. 4258, pp. 36–58.