An Overview of Mathematical Models for Data Privacy

Tamara Stefanović[1], Silvia Ghilezan [2,3]

[1,2] *University of Novi Sad*

[3] *Mathematical Institute SASA*

E-mail: [1] tstefanovic@uns.ac.rs, [2] gsilvia@uns.ac.rs

For centuries, people have shared information with each other and with institutions. In the last few decades the development of technology has made possible to manipulate a large amount of data, but at the same time it has developed data privacy problems. The problem of data privacy concerns how data is collected and stored, whether and how data is shared with a third party, as well as which laws are governing data sharing in areas such as health care, education and financial services ([6]). We give an overview of two fundamental mathematical models for describing data privacy problems that significantly differ from the traditional privacy approach - privacy in context ([1]) and differential privacy ([2]).

Privacy in context. Contextual integrity represents a philosophical account of privacy in terms of transfer of personal information. Here the term personal information" refers to any information related to an identified or identifiable natural person, as Helen Nissenbaum defined in [4]. A formal framework for expressing norms of transmission of personal information, inspired by contextual integrity, was presented in [1]. A temporal logic is used to capture the principles of information transmission. Formulas are generated by the following grammar:

$$\varphi ::= send(p_1, p_2, m) | contains(m, q, t) | inrole(p, r) |$$

$$incontext(p, c) | t \in t' | \varphi \wedge \varphi | \neg\varphi | \varphi \mathcal{U} \varphi | \varphi \mathcal{S} \varphi$$

$$\bigcirc\varphi | \exists x : \tau . \varphi .$$

Information about a subject is transmitted through a communication action from a sender to a recipient:

- $send(p_1, p_2, m)$ holds if agent p_1 sent the message m to agent p_2
- $contains(m, q, t)$ holds if message m contains the attribute t of agent q.

For simplification, it is assumed that information describes a single individual. However, the model includes computation rules enabling communicating agents to combine messages to compute additional information: $t \in t'$ holds

if attribute t can be computed from attribute t'. Communicating agents are associated with roles as a part of contexts, and depending on the role, communication can be permitted or prohibited:

- $inrole(p, r)$ holds if agent p is active in role r
- $incontext(p, c)$ holds if agent p is active in a role of context c.

This model is convenient for formalizing privacy laws because each privacy law is drawn to protect certain types of information in particular contexts, such as health care, employment, the marketplace and so on. Up to now, it has been used to formalize several privacy laws, such as GLBA (Gramm-Leach-Bliley Act), HIPAA (Health Insurance Portability and Accountability Act) and COPPA (Children's Online Privacy Protection Act).

Differential privacy. Privacy can also be considered from the perspective of statistical analysis of data or the release of statistics derived from personal data. Suppose a trusted curator is managing a sensitive database and needs to release some statistics from this data to the public. Also suppose there is an adversary who wants to reveal or to learn some of the sensitive data. Differential privacy ([2]) proposed by Cynthia Dwork relies on incorporating random noise so that everything an adversary receives is noisy and imprecise. The question is what kind of random noise to use so that the results still can be useful. The main challenge is achieving privacy while minimising the utility loss.

Let $D \in \mathcal{D}^n$ be a database. A query q is a function applied on a database D ([2]). We say $\mathcal{M}$ is a *privacy mechanism* or simply *mechanism* obtained by adding noise if for every query q, $\mathcal{M}$ creates a new randomized query $q^*(D) = q(D) + noise$. Let $D, D' \in \mathcal{D}^n$ be two databases that differ in at most one entry, we call them *adjacent databases*.

Definiton. Let $\varepsilon > 0$. A mechanism $\mathcal{M}$ is ε -differentially private iff for every pair of adjacent databases D, D' and for every $S \subseteq range(\mathcal{M})$:

$$Pr[\mathcal{M}(D) \in S] \leq exp(\varepsilon) Pr[\mathcal{M}(D') \in S],$$

where the probability space is over the coin flips of the mechanism $\mathcal{M}$.

The following example shows what differential privacy actually provides. Suppose Alice obtains database $D \in \mathcal{D}^n$ with n entries. She provides Bob with output o of a mechanism $\mathcal{M}(D)$. Bob knows the values of $n-1$ entries (database D_1), and has to guess the value of the n-th entry (d_n). For each possible value x of d_n, Bob can learn the distribution induced by $\mathcal{M}(D_1 \cup \{x\})$ and then pick x assigned to highest probability of the output. But, if $\mathcal{M}$ is ε -differentially private, for every $x, y \in \mathcal{D}$ holds

$$Pr[\mathcal{M}(D_1 \cup \{x\}) = o] - Pr[\mathcal{M}(D_1 \cup \{y\}) = o] \leq \varepsilon.$$

Therefore, Bob cannot do better than random guessing. In fact, if an individual is considering to allow her/his data to be used or not, by the promise of differential privacy, she/he can be almost indifferent between these two choices, because participating will not cause any additional harm.

Differential privacy has also been used to formalize privacy laws, for example FERPA (Family Educational Rights and Privacy Act), but the best known users of differential privacy models are certainly Apple and Google.

A brief comparison of the methods. The contextual integrity framework considers privacy from the perspective of information flow and uses temporal logic formulas to model privacy norms. On the other hand, differential privacy considers privacy from the perspective of statistical analysis and releasing statistics of personal data. The fundamental difference between these two approaches is in underlying mathematical methods: logic and probability. Also, formal logical model may allow sharing some personal information depending on agents role, for example: a doctor can share patients private medical information with that patient. What is not included in the formal logical model is the communication about aggregate statistics. For example, communication restriction such as the average salary of bank managers can be released only if it does not identify a particular individuals salary" cannot be expressed in formal logical model, but it is precisely the type of restriction expressed in the differential privacy model ([5]).

Future work. We are currently exploring the possibilities of combining different kinds of privacy formalization including inverse privacy ([3]).

Acknowledgment. This work has been partially supported by the Science Fund of the Republic of Serbia under grant AI4TrustBC (6526707).

References

[1] A. Barth, A. Datta, J. C. Mitchell and H. Nissenbaum, Privacy and Contextual Integrity: Framework and Applications, in *Symposium on Security and Privacy*, (Berkeley, California), pp. 184-198, IEEE, Computer Society, 2006.

[2] C. Dwork and A. Roth, The Algorithmic Foundations of Differential Privacy, *Foundations and Trends in Theoretical Computer Science*, vol. 9, pp. 211-407, 2014.

[3] Y. Gurevich,E. Hudis and J. M. Wing, Inverse Privacy, *CoRR*, vol. 1510.03311, 2015.

[4] H. Nissenbaum , *Privacy in Context - Technology, Policy, and the Integrity of Social Life*, Stanford University Press, 2010.

[5] K. Nissim, A. Bembenek, A. Wood, M. Bun, M. Gaboardi, U. Gasser, D. R. O'Brien, T. Steinke and S. Vadhan Bridging the Gap between Computer Science and Legal Approaches to Privacy,*Harvard Journal of Law & Technology*,vol. 31, pp. 689-713, 2018.

[6] D. J. Solove, A taxonomy of privacy, and the Integrity of Social Life, *University of Pennsylvania Law Review*, vol. 154, pp. 477-560, 2010.