Technical Perspective:
Graph Theory for Data Privacy: A New Approach for Complex Data Flows

Elena Ferrari
Department of Theoretical and Applied Sciece
University of Insubria, Varese (Italy)
elena.ferrari@uninsubria.it

Nearly all of the world's population now uses online services that request personal information, covering almost every aspect of our lives. The abundance of personal data in digital form has brought incredible benefits to end users, enabling them to access personalized and advanced services based on the analysis of the data collected. This capability has dramatically improved the user experience in various application domains, ranging from healthcare to e-commerce, finance, logistics, and entertainment, to name a few. Numerous technological advancements in the field of big data have enabled this massive processing of personal data, and recent advances in AI data processing capabilities will expand the ways in which service providers will use personal data in the coming years. Machine learning algorithms, powered by AI, will be used to make increasingly accurate predictions about user behavior by uncovering hidden correlations within massive data sets. There is therefore a tension between the desire to fully exploit personal data in such ecosystems and the need to provide strong privacy and transparency guarantees to the individuals whose data is being exploited. Privacy protection is further complicated because data processing is typically not performed in isolation but through pipelines of different services, with each step making inferences about the personal data consumed by the services in subsequent steps.

Privacy and transparency in data processing have also been driven by the emergence of privacy regulations, the most notable being the European General Data Protection Regulation (GDPR), which has inspired many subsequent privacy laws, such as the California Consumer Privacy Act (CCPA), or the Brazilian Lei Geral de Proteção de Dados (LGPD). A pillar of the GDPR is that end users should have control over how their data is used, and for what purposes the data is managed and processed.

As a result, the field of privacy-preserving techniques has been a very active research area in the last decades, and many privacy-enhancing techniques have been proposed. However, the majority of these techniques protect privacy by masking or perturbing data, the main notable ones being those based on differential privacy, which is today a gold standard of data privacy. However, how to strike a balance between utility and privacy achieved through data modification is still an open issue in many use-case scenarios.

In contrast, the development of solutions to help end users in controlling the flow of their personal data when data are processed by companies' data processing workflows is still an open research issue. The main challenge is how to be compliant with user privacy constraints by, at the same time maximizing data utility for the service provider. A further relevant dimension of the problem is that these processing workflows often target massive processing of personal information. For instance, the highlighted paper reports that the data processing workflow of Meta contains over 12 million service instances and over 180,000 communication edges between services. Developing a sound and scalable solution for this setting is far from being trivial.

The highlighted paper addresses this challenging scenario by proposing a theoretically sound and effective solution. The key intuition of the paper is the use of a graph model: data processing is modelled as a graph where, in addition to nodes representing the types of data collected and the processing algorithms, there is another class of nodes denoting the purposes of data processing. Purpose nodes are linked to an associated utility. In this way, privacy constraints can be modeled as pairs of nodes in the graph that should be separated in order to satisfy the constraints. The problem of satisfying users' privacy constraints is then expressed as a graph optimization problem: selecting the nodes in the graph to separate in a way that maximizes utility. The authors show that this problem is NP-hard and present several alternative heuristics and related algorithms, precisely characterizing their underlying computational complexity as well as the accuracy achieved.

The authors then provide a comprehensive performance evaluation for a relevant instance of the defined optimization problem targeting linearly additive utility functions. Experiments on different datasets show that the designed algorithms can achieve a near-optimal solution in a few seconds for scenarios with graphs of thousands of nodes and tens of user constraints.

The paper not only presents a solid and practical solution to a relevant and timely problem in the area of privacy-preserving data management, but also opens the way for many follow-up research contributions, some of the most important of which are discussed at the end of the paper.