SEmantic Networks of Data: Utility and Privacy

Cédric EICHLER*†, Pascal BERTHOMÉ*†, Jacques CHABIN*‡, Rachid ECHAHED§,
Mirian H. FERRARI*‡, Benjamin NGUYEN*†, Frédéric PROST§
*Laboratoire d'Informatique Fondamentale d'Orléans
†INSA Centre Val de Loire, Email: firstname.lastname@insa-cvl.fr
‡Université d'Orléans, Email: firstname.lastname@univ-orleans.fr
§Laboratoire d'Informatique de Grenoble, Université Grenoble Alpes,
Email: firstname.lastname@univ-grenoble-alpes.fr

I. Introduction, practical information

The SEmantic Networks of Data: Utility and Privacy (SEND UP) project aims at ensuring privacy, anonymity, and usefulness in (open) linked data. To this end, SEND UP brings together the LIG (Laboratoire d'Informatique de Grenoble) and LIFO (Laboratoire d'Informatique Fondamentale d'Orléans) laboratories. It started in November 2018 under Cédric Eichler's coordination and is expected to end in October 2022. SEND UP is founded by the ANR under the JCJC (young researcher) funding instrument with the reference ANR-18-CE23-0010 following ANR's 2018 Generic Call for Proposals (AAPG 2018).

II. Context and objectives

The amount of data produced by individuals and corporations has dramatically increased during the last decades. This generalized gathering of data brings opportunities (e.g., building new knowledge using this "Big Data") but also new privacy challenges. The general public express a growing distrust over personal data exploitation, which has been met with successive strengthened regulations (e.g. EU general data protection regulation, GDPR). In the meantime, open data is taking a crucial place within many administrations. The open data policy is a powerful move by public institutions aiming at publishing data collected by public agent. The objective is to manage this data as an asset to make it available, discoverable, and usable by anyone. This leads to an important new societal challenge at the crossroads of these social evolutions: how can privacy be preserved while publishing useful data?

This challenge has led to a growing interest for data sanitization, the art of disclosing personal data without jeopardizing privacy, and data-set anonymisation. An anonymized dataset is a dataset which is difficult, costly, or impossible to relate to real individuals.

Nowadays, data are often organized as graphs with an underlying semantic to allow efficient querying and support inference engines. Such is the case in, for example, linked data and semantic web typically relying on RDF. The SEND UP project focuses on such databases and will follow two main goals: (1) prevent illegitimate use of private data while querying semantic data graphs and (2) publish useful sensitive semantic data graphs will preserving privacy.

III. Scientific barriers

A massive amount of work has focused on privacy in data presented as tables, resulting in multiple well-established models, such as k-anonymity, l-diversity, and differential privacy. More recently, these concepts have been translated and applied to graph representations, but mainly in the context of social networks. These methods usually consider homogeneous nodes with no semantic relation and aim at protecting the graph topology. More often than not, their utility is experimentally evaluated with regard to specific sets of functions and/or graph characteristics (e.g., diameter, max degree and degree distribution...). To achieve semantic data graph sanitization, the SEND UP project aims at:

- Introduce knowledge-based and usage-based utility metrics, related to facts present in, or that can be deduced from, the base. Indeed, due to the nature of the targeted graph utility metrics and evaluation can not rely on the preservation of, for example, the diameter of the graph.

- Fully define the side-effects of transformations in semantic graph databases and introduce methods and tools to handle them. Indeed, updating instances of semantic data graphs during their sanitization implies many new difficulties including side-effects on the instances but also on their schema and constraints. The sanitization context brings issues that have been mildly studied in the literature (e.g., updating incomplete data-bases, triggering schema/constraints evolutions as side-effects of instance updates...) and even completely new ones (e.g., solving non-deterministic updates as an optimization problem regarding privacy and utility metrics).

- Introduce new sanitization concepts granting privacy guarantees in semantic graph databases and taking into account vertex heterogeneity and the existence of logical relations and semantic rules between attributes.

- Introduce methods and algorithms for semantic graph databases sanitization integrating new expanded anonymity concepts, usage-based and knowledge-based utility metrics but also transformations side-effects. Efficient techniques should account for side-effects during the decision process rather than merely triggering them afterward.

These objectives are to be supported by a suite of software modules validated in lab (TRL 4 - technology validated in lab) implementing our proposed metrics and algorithms.