SEmantic Networks of Data: Utility and Privacy

Cédric EICHLER*† Jacques CHABIN*‡, Rachid ECHAHED§,
Mirian H. FERRARI*‡, Nicolas HIOT*‡, 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[1] 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 April 2023. 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). SEND UP was first presented at RESSI 2019 as a starting project.

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.

[1] www.univ-orleans.fr/lifo/evenements/sendup-project

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:

1- 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.

2- 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).

3- 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.

4- 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.

IV. Realizations

At the time being, contributions have been proposed in two main axis that can be roughly assimilate to points 2 and 3.

A. Update management in RDF/S database

We introduced SetUp [2], [3], a theoretical and applied framework for the management of RDF/S database evolution under on the basis of *graph rewriting rules*. In SetUp, we consider databases satisfying integrity constraints –a well known variant of RDF/S semantic– and the Closed World Assumption (CWA) semantics. Indeed, we argue that the OWA is not adapted to data centric applications needing complete and valid knowledge and to data sanitization in particular.

Each atomic update is formalized using a graph rewriting rule whose application *necessarily* preserves the constraints. Furthermore, SetUp manages updates by ensuring rule applicability through the generation of side-effects: new updates which guarantee that rule application conditions hold.

The first version of the tool solved determinism by applying arbitrary choices. Furthermore, preliminary evaluation of SetUp [2] showed an explosion of generated side-effects in some specific scenariis implying schema updates.

In [1] we proposed two major improvements of SetUp , $\mathsf{SetUp}_{\mathsf{opt}}$ and $\mathsf{SetUp}^{ND}_{\mathsf{opt}}$ concerning the management ofside-effects. $\mathsf{SetUp}_{\mathsf{opt}}$ avoids the generation of unnecessary verifications while recursively handling side-effects. This is done by leveraging knowledge of updates' history. Indeed, while we do not have any a priori knowledge on an original update U, we do have some new knowledge when dealing with U's side-effects. $\mathsf{SetUp}^{ND}_{\mathsf{opt}}$ is a module of $\mathsf{SetUp}_{\mathsf{opt}}$ proposing a flexible and modular way of handling non-determinism. $\mathsf{SetUp}^{ND}_{\mathsf{opt}}$ generates all different ways of guaranteeing the applicability of an update. The corresponding set of ordered lists of side-effects is transmitted to an easily customizable choice function. This function selects one of these lists and returns it to $\mathsf{SetUp}^{ND}_{\mathsf{opt}}$ which ensures its recursive application. To demonstrate modularity and evolution capability, we implemented three different choice functions, including one taking user's input through a TUI.

Experimental evaluation showed that the combination of these both improvements successfully handle the aforementioned problematic secnarii. We also demonstrated $\mathsf{SetUp}_{\mathsf{opt}}$ś correction: it does terminate and do apply the required update (if the user has sufficient rights and the update is not intrinsically inconsistent).

A demonstration of SetUp can be proposed during RESSI.

B. Using projection to enable differential privacy over RDF datasets

Differential privacy (DP) is currently one of the most popular and prevalent definitions of privacy. It relies on releasing query results in such a way that observing a result does not provide much information on whether a particular database or one of its neighbour has been queried. To achieve DP in semantic directed graphs, we addressed two main issues: 1) what it means to be a neighbouring dataset and 2) reducing sensitivity. In directed graph hover, many queries are highly sensitive to small modifications of the original graph, which means directly that directly perturbing the results to obtain DP is not a good option (or not even possible). In [**?**], we propose a new approach based on graph projection to adapt differential privacy to edge-labeled directed graphs, i.e. RDF graphs, while reducing the sensitivity of various queries. The idea is to project the graph on a subspace where a query do not vary much over a neighbourhood. We provide an analysis of several variant of this approach depending on the neighbourhood/distance notion and related privacy model: node-DP, edge-DP, and typed-edge-DP. We demonstrated that most projections preserve neighbourhood when associated with the relevant distance. Therefore, the sensitivity of the composed mechanism (project $\circ$ query) is exactly equal to the sensitivity of the query over the projected space.

We are currently working on further analytical and experimental evaluation of the approach as well as optimizing the projection process to minimize data loss. More interestingly, we wish to associate this approach with the work on RDF evolution under constraint. We plan to investigate the theoretical and practical impact of considering constraints and *invalid* dataset on the previous approach and the DP properties. Indeed, a DP mechanism over the set of all RDF graphs could be vulnerable to privacy leakage if applied to the set of *valid* RDF graphs –e.g. if some database has no valid neighbour–.

References

[1] Jacques Chabin, Cédric Eichler, Mírian Halfeld Ferrari, and Nicolas Hiot. Graph rewriting rules for RDF database evolution: optimizing side-effect processing. *Int. J. Web Inf. Syst.*, 17(6):622–644, 2021.

[2] Jacques Chabin, Cédric Eichler, Mirian Halfeld-Ferrari, and Nicolas Hiot. Graph rewriting rules for rdf database evolution management. In *Proceedings of the 22nd International Conference on Information Integration and Web-Based Applications & Services*, iiWAS '20, page 134143, New York, NY, USA, 2020. Association for Computing Machinery.

[3] Jacques Chabin, Cdric Eichler, Mirian Halfed Ferrari, and Nicolas Hiot. SetUp: a tool for consistent updates of rdf knowledge graphs.

[4] Sara Taki, Cédric Eichler, and Nguyen Benjamin. Using projection to improve differential privacy on rdf graphs. In *Confrence sur la Gestion de Donnes Principes, Technologies et Applications*, BDA 2021, 2021.