SEmantic Networks of Data: Utility and Privacy

Cédric Eichler, Pascal Berthomé, Jacques Chabin, Rachid Echahed, Mirian Halfeld-Ferrari, Benjamin Nguyen, Frederic Prost

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. 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).

This 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. Both domains aim to maintain a certain data quality in order to produce information as useful as possible.

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 SEmantic Networks of Data: Utility and Privacy (SEND UP) project started in November 2018 and will look for more efficient sanitization and anonymization techniques for data stored as graphs with an underlying semantic. In this regard, SENDUP integrates two general goals.

A. Querying sensitive data

An individual or company may have legitimate motivations to query a database it does not own, but preventing illegitimate access to private information while allowing legitimate database usage is a difficult task. In general, the objective is to limit the knowledge a tenant may gather from the database by limiting the queries it may execute. This is particularly difficult when the database possesses an underlying semantic, as attributes may have strong correlations and since inference rules can generally be exploited to infer new knowledge.

B. Publishing open sanitized semantic data

Open data is taking a crucial place within many administrations and industries. The objective is to manage data as an asset to make it available, discoverable, and usable by anyone. However, publishing data must not harm the privacy of the citizens. In linked data, qualifying the auxiliary knowledge of a potential attacker and the knowledge it can produce from a sanitized database is especially challenging. Hence, certified sanitized techniques are of utmost importance in this respect.

II. Scientific barriers and expected results

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. If some recent works target graph databases, graph anonymity is mainly considered 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...).

A. Utility metrics

Due to the nature of the targeted graph, utility metrics should differ from existing works. Indeed, depending on the expected usage, it could make little to no sense to preserve, for example, the diameter or the degree distribution of the graph. Thus, this projects aims at introduce **appropriate knowledge-based and usage-based utility metrics for graph databases** related to facts present in, or that can be deduced from, the base. A particular challenge is to provide relevant yet generic (i.e. non add-hoc) utility metrics.

B. Sanitization and privacy guarantees

Anonymity methods for regular graphs cannot be applied in this context. In fact, nodes in semantic networks are heterogeneous, contrary to, for example, online social networks where nodes represent users. They should be handled differently depending on their nature (i.e. whether they represent individuals or other data, identifying, pseudo- or non-identifying data, sensitive or non-sensitive data...). Secondly, attributes in such graphs often have logical relations and can not be considered as independent, which is almost always the case in existing sanitization techniques. Furthermore, as discussed in the next paragraph, semantic data-graphs are subject to inherent constraints that should be taken into account in the sanitization process. In addition to expanding graph anonymity concepts to semantic data graphs, the proposed methods

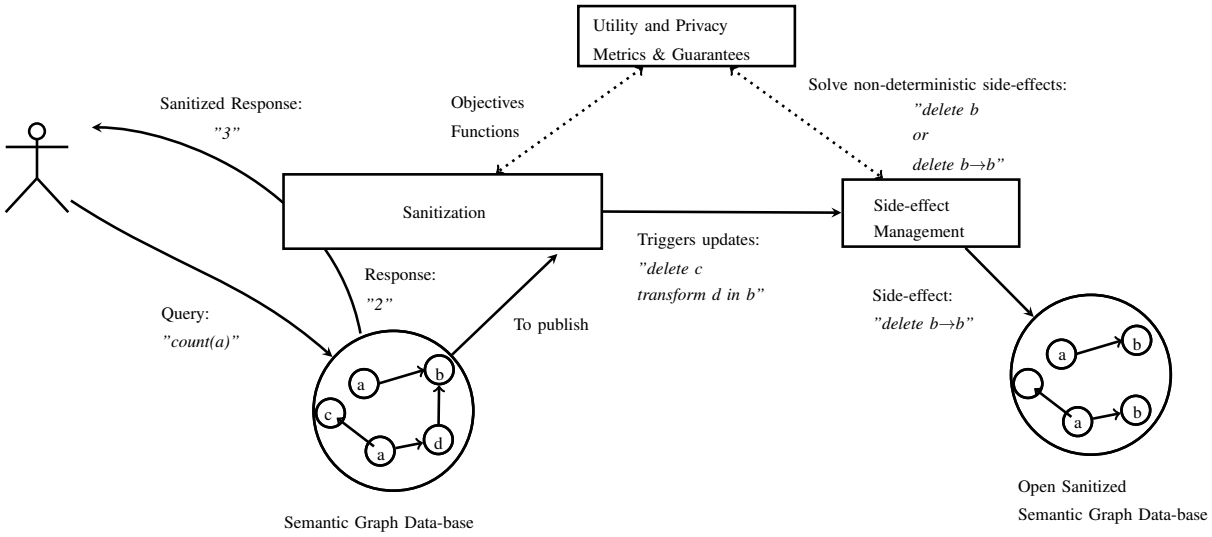

Figure 1. A simplified view of SEND UP's targeted scenarios and software

should be accompanied by formal privacy guarantees such as differential privacy. Accordingly, SENDUP will introduce **new sanitization concepts granting privacy guarantees in semantic graph databases**.

C. Updating semantic data graphs

Graph modifications in semantic networks are subject to new constraints. For instance, in the semantic web, RDF resources are usually associated with a set of structuring constraints that are typically defined in RDFS or, more recently, in ShEx or SHACL. Integrity constraints may also be imposed in order to ensure data quality. Accordingly, these different constraints must be maintained by the sanitization process. Moreover, if the sanitization processes may induce side-effects within the anonymized graph, they can go even beyond: schema (or even integrity) constraint evolution may be activated. Mastering these side-effects and taking them into account during the sanitization process is crucial to guarantee an acceptable utility of the sanitized data. In spite of existing work, fully mastering side-effects is in itself a scientific challenge. In this project we focus on **mastering updates' side-effects and their interferences with the graph sanitization process**. We seek to minimize knowledge loss and maximize utility when guiding transformations in the sanitization process in spite of their side-effects.

D. Providing re-usable tools

Finally, the SENDUP project will provide a suite of software modules implementing our proposed utility metrics and sanitization algorithms integrating updates' side-effects. We target the European technology readiness level TRL 4 - technology validated in lab. A simplified view of software modules, their high level interactions and two scenarios related to the two goals described in section I are represented Fig. I-B.

The first scenario consists in a user querying an undisclosed database. The response will be curated by the sanitization module (M1) according to utility and privacy metrics provided by a dedicated module (M2), as well as an history of previous requests. This process should also integrates constraints of the data-bases as well as associated inference rules.
In the second scenario, a database is sanitized to be openly published. M1 updates the database to meet privacy guarantees provided by M2, which can lead to side-effects. In the example, we suppose that (i) due to some structuring constraint, nodes labelled "b" can not be neighbours (ii) following the updates triggered during the sanitization, two nodes "b" are neighbours. Correcting side-effects are thus triggered by a dedicated module (M3). Here, M3 could either trigger the deletion of one of the two nodes "b" or the edge $b \rightarrow b$. The side-effect providing the best privacy/utility trade-off as evaluated by the metrics module is therefore selected and applied.

This is obviously a very simplistic example. In fact, side-effects may trigger side-effects. Furthermore, M1 could also have several initial update possibilities. Every decision should anticipate its consequences and pick the best trade-off. Each module will itself follow a modular implementation or at least propose several implementations (as several metrics and several sanitization techniques will be proposed).

III. Project description, administrative information

The SEND UP project 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).