SEmantic Networks of Data: Utility and Privacy

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

Project

SEmantic Networks of Data: Utility and Privacy

Laboratoire d'Informatique Fondamentale d'Orléans,
Laboratoire d'Informatique de Grenoble
Coordinator contact: `cedric.eichler@insa-cvl.fr`

Abstract

The amount of data produced by individuals and corporations has dramatically increased during the last decades. This generalized gathering of data brings opportunities but also new privacy challenges. 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 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 while preserving privacy.

Context and goals

More and more private information collected: threatening privacy yet useful ⇒ needs for privacy guarantees and utility preservation.

Two main **scenarios:**

- Publishing a *useful* anonymised data-base
- *Accurately* answering queries without jeopardizing privacy

Target data-base = graph data-base with an underlying semantic

- used in linked data, semantic web...
- e.g. RDF

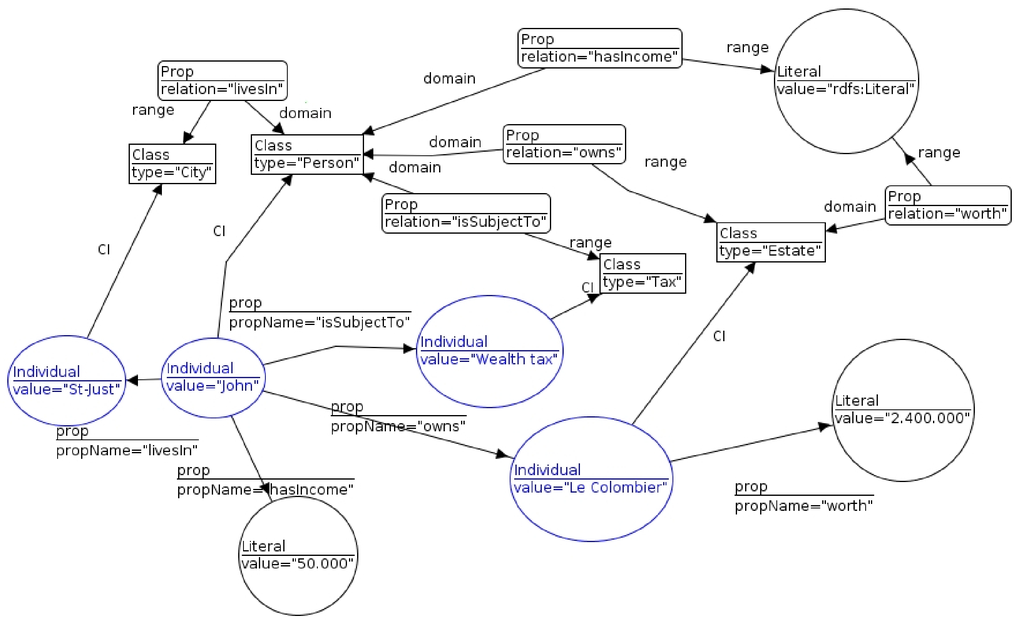

Figure 1: Example of a semantic graph DB

Graph & privacy

Multiple established models on "classical" data-bases

- k-anonymity, l-diversity,...
- differential privacy

Some "recent" extensions to graphs

- Mostly homogeneous nodes with no semantic (e.g., OSN [1])
- Usually aims to protect topology

Graph privacy techniques

- Differential privacy: how to qualify "neighbouring" data-set?
  - Usually edge-DP
  - Rarely node-DP (e.g. [2] regarding degrees)
  - ▷ **Target privacy technique:** More than node-DP, person-node-DP
- k-anonymity: k what?
  - Usually k-degree anonymity (e.g. [3])
  - ▷ **Target privacy technique:** k-pattern anonymity
- ▷ Issue: semantic and inter-dependant data

Evaluating utility

Usually:

- Related to topology preservation (e.g. diameter, degree distribution)
- Minimal transformation
- ▷ Issue: still semantic and inter-relations!
- Possibility: add-hoc or user defined metrics (e.g. [5])
- ▷ **Target utility metrics:** Knowledge and usage-based

Updating semantic data graphs

Anonymizing a graph → graph instance updates. Not that easy!

- Incomplete information?
- Structuring (e.g. RDF/S, ShEx) and integrity constraints:
  - Forbidding updates?
  - ▷ **Target constraints management:** Triggering instance side-effects [4]
    - Non-determinism
    - Impact of the initial update on the utility?
    - Should allow schema/constraints updates?

Targeted software and scenarios

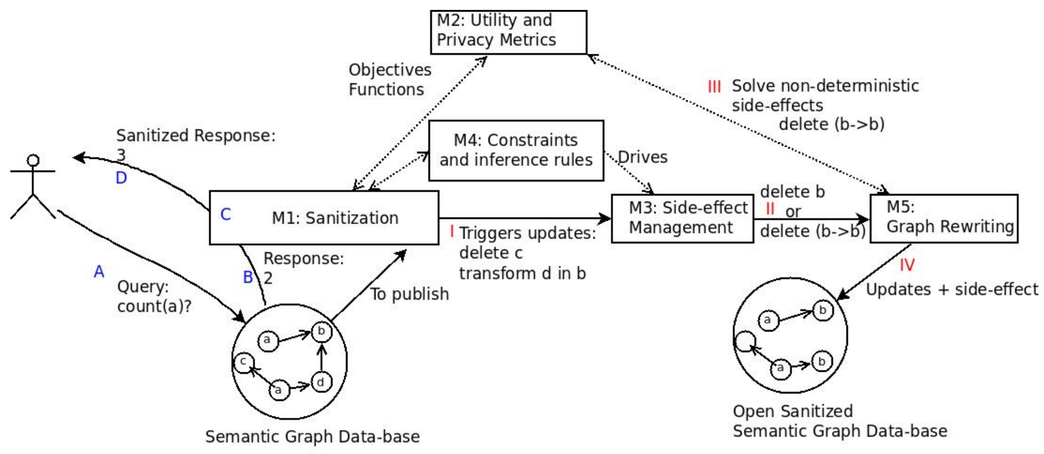

Figure 2: Targeted architecture

Sanitization of query on undisclosed data:

A. A user queries the data-base.
B. The "Sanitization" module (M1) gets the result.
C. M1 perturbs the results according to the "Privacy and utility metrics" module (M2), with constraints provided by a dedicated module (M4).
D. The user gets a curated answer.

Publishing anonymised graph data-bases:

I. M1 updates the database to meet privacy guarantees provided by M2 (e.g. *delete c* and *transform d into b*)
II. Driven by M4, M3 infers possible side-effects (e.g. *delete b* or *delete edge b→b*)
III. The set of side effects with the best utility/privacy trade-off as evaluated by M3 is picked (e.g. *delete edge b→ b*)
IV. A graph transformation module (M5) applies the selected modifications and the anonymised data-base is published.

Project information

- Partners and expertise:
  - LIFO, INSA Centre Val de Loire → privacy
  - LIFO, Université d'Orléans → updates on RDF databases
  - LIG, Université Rhône Alpes → graph rewriting and transformation
- Start/end date: Nov. 2018/2022
- Funding: ANR JCJC

[1] E. Zheleva and L. Getoor. *Privacy in social networks: A survey*. In Social network data analytics,pages 277–306. Springer, 2011.
[2] W.Y. Day, N. Li, and M. Lyu. *Publishing graph degree distribution with node differential privacy*. In Proceedings of SIGMOD '16, pages 123– 138. ACM, 2016.
[3] T. Tassa and D. J. Cohen. *Anonymization of centralized and distributed social networks by sequential clustering*. In IEEE TKDE, 25(2):311–324, 2013.
[4] M. H. Ferrari and D. Laurent. *Updating RDF/S databases under constraints*. In Proceedings of Advances in Databases and Information Systems, pages 357–371, 2017.
[5] R. Delanaux, A. Bonifati, M-C. Rousset, and R. Thion. *Query-Based Linked Data Anonymization.* In Proceedings of APVP'18, 2018.

C. Eichler, P. Berthomé, J. Chabin, R. Echahed, Mirian H. Ferrari, B. Nguyen, F. Prost. *Project SEmantic Networks of Data: Utility and Privacy*. APVP, 9-11 june 2019