Semantic attack on graph databases

Project: This internship is part of project Semantic Networks of Data: Utility and Privacy (SENDUP[1]) that studies data privacy in graph databases (e.g., RDF) with underlying semantics.
Laboratory & Team: LIFO, Systems and Data Security team, INSA Centre Val de Loire, 88 boulevard Lahitolle 18022 Bourges
Duration and start: 5 to 6 months, at the candidate's earliest convenience.
Contact: to apply or request additional information, send a mail and a resume to adrien.boiret@insa-cvl.fr
Requirements:

- Master, Bac +5 in computer science / engineering or equivalent
- Knowledge or interest about databases (especially graph databases, e.g. RDF) and data privacy
- Ability to read and write english documents
- Proficiency in a coding language (preference for Java)
- Willingness to work in autonomy and in a team

Subject:

Data safety and privacy are concerns currently receiving intense attention, notably through the introduction of GDPR reglementations that aim to ensure data collection, treatment, and publication never trespass on a person's right to privacy. The notion of differential privacy (DP) grew popular as a yardstick of privacy for data publication processes, where a database containing sensitive information can still answer queries without compromising privacy.

The guaranty provided by DP is that it is difficult to differentiate between a graph and one of its neighbours (i.e. the same graph differing on exactly one information) when observing the answer to a query. This is a convincing guaranty of privacy, as it means that a graph yields results so similar to its neighbours', that an attacker cannot deduce with certainty any specific information in a graph. However, this guaranty works best under the assumption that any graph has neighbours to "hide behind". If a graph is isolated from any of its neighbours, then the guaranty provided by DP weakens.

We posit that such situations can arise if the graph databases we consider are known to follow structural constraints (e.g. "every patient has a doctor") or semantic constraints (e.g. "Dr Wilson is an oncologist"). If all possible graphs must follow specific rules, then it is possible that some graphs have no neighbours that an attacker could confuse them with.

In this internship, we aim to formalise and evaluate through experimentation the damage that prior knowledge of a target graph's schema can make on the privacy of a DP-guarantying process.

Goals and Objectives:

- Identify an example of an attack through schema knowledge of a DP-guarantying process

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

- Evaluate the risk and impact of such attacks on current privacy standards
- Study mechanisms, such as metric-based d-differential privacy[2], as a mean to counter such attacks

Keywords: Data privacy, Differential privacy, Databases, Graph databases, Graph ontology, Bayesian statistics

[2] https://hal.inria.fr/hal-00767210/document