Data Privacy on Graphs with semantic information

The Systems and Data Security Team of Laboratoire d'Informatique Fondamentale d'Orléans is offering a PhD position to study data privacy in graph databases containing semantic information. The goal is to explore how privacy guaranties potentially weaken in the case where graph databases respect a known schema or ontology, and to present adaptations and countermeasures.
Where: The PhD will take place at LIFO's INSA Centre Val de Loire campus, located in Bourges.

INSA Centre Val de Loire in Bourges is a human sized establishment (approx. 800 students, including approx. 200 in cybersecurity). The Bourges campus has affordable student housing and restauration. For non-French speakers, French classes are made free and available. Partial remote work and flexible working hours can be organized.

The Systems and Data Security team (SDS) is an active and international team of approx. 20 people working in various topics concerning computer security and data privacy in the broad sense.
Financing:

- This PhD position is part of the CyberINSA project
- Gross salary is 2050€ per months before taxes.
- Complementary remuneration for additionnal teaching is negotiable.

Requirements:

- Research Master in computer science / engineering
- 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

Contact and Application: The PhD thesis will be directed by Benjamin Nguyen (benjamin.nguyen@insa-cvl.fr), and will be co-supervised by Adrien Boiret (adrien.boiret@insa-cvl.fr). To apply or request additional information, send a mail with contact information, a resume, and a motivation letter to adrien.boiret@insa-cvl.fr.

Subject:

The quantity of personnal data online increased dramatically in the past decade. This represents new opportunities but also a great number of pressing questions about preserving the safety and privacy of sensitive data. These questions receive 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 general goal of privacy is to allow limited access to a database that has a use for an honest stakeholder while preventing attackers from deducing specific personal information considered sensitive. This can entail the publication of the data once it has been sanitized, i.e. every identifiying or sensitive information has been removed or obfuscated as to prevent an attacker from making undesirable inferences. This can also be achieved by allowing users to ask general queries that do not expose specific informations, usually adding some noise to the answer.

The notion of differential privacy [3, 4] (DP) grew popular as a yardstick for data sharing 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. two graph databases that only differ in the value – or presence – of one specific datum within) 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 that knows everything about the original graph except for a single information cannot deduce it with certainty from the process's output.

However, the guaranty provided by DP 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.

The aim of this PhD position is to study how DP and data privacy evolves in semantic databases [6, 2] when they are known to follow such constraints, for instance as expressed in an ontology (RDFS [5] or OWL [1]). We want to identify, detect and quantify the possible loss of privacy incurred, detail how malicious users could exploit this weakness through a semantic-based attack of a DP process, then establish how best to adapt DP (and methods to provide it) to propose countermeasures that ensure privacy is preserved without destroying the querying processes' usefulness in the process.

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

References

[1] OWL 2 web ontology language document overview (second edition). W3C recommendation, W3C, December 2012. https://www.w3.org/TR/2012/REC-owl2-overview-20121211/.

[2] Tim Berners-Lee. Relational databases on the semantic web, 1998. Technical report, W3C, 2015. https://www.w3.org/DesignIssues/RDBRDF.html.

[3] Cynthia Dwork. Differential privacy. In *ICALP (2)*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2006.

[4] Cynthia Dwork. Differential privacy: A survey of results. In *TAMC*, volume 4978 of *Lecture Notes in Computer Science*, pages 1–19. Springer, 2008.

[5] Ramanathan Guha and Dan Brickley. RDF schema 1.1. W3C recommendation, W3C, February 2014. https://www.w3.org/TR/2014/REC-rdf-schema-20140225/.

[6] Ora Lassila. Resource description framework (RDF) model and syntax specification. W3C recommendation, W3C, February 1999. https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.