Data Privacy for Graphs with Semantic Informations

Yasmine Hayder, Adrien Boiret, Cédric Eichler, Benjamin Nguyen

Data Privacy for Graphs with Semantic Informations

Yasmine Hayder, Adrien Boiret, Cédric Eichler and Benjamin Nguyen
INSA Centre Val de Loire, Inria
FirstName.LastName@insa-cvl.fr

ABSTRACT

This PhD focuses on the privacy implications of querying graph data structured in Resource Description Framework RDF and governed by Web Ontology Language OWL semantics. We have identified privacy breaches in these graph structures when queried. The aim of this thesis is to address and remedy these issues.

KEYWORDS

Data Privacy, OWL , RDF , Differential Privacy.

1 INTRODUCTION

Preserving privacy in graph data has become a critical concern. RDF graphs, which follow OWL constraints, present unique challenges in maintaining data privacy. Despite existing privacy protection concepts, these methods often fall short when applied to ontology-based graphs. The following work presents an experiment that demonstrates the vulnerabilities in current privacy techniques.

2 MOTIVATIONAL EXPERIMENT

To illustrate the generality of this problem and to provide realistic, easily demonstrable scenario, we have selected this particular motivational example. This choice aims to highlight how adding OWL constraints to the graph can potentially increase the risk of privacy breaches.

Consider a family tree, depicted in figure 1, representing mitochondrial disease, where each affected mother has all her children affected, and each affected father has none of his children affected. Consequently, the "parentof" relation is inherently sensitive. We aim to protect such relations.

Mitochondrial disease :

$(A \text{ isWomen}) \wedge (A \text{ isAffected}) \wedge (A \text{ ParentOf } B) \rightarrow (B \text{ isAffected})$

The graph represents a family linked solely by the "parentof" relation. "Childof" and "siblings" relations are inferred based on the following OWL axioms:

A ParentOf $B \rightarrow \neg(B \text{ ParentOf } A)$: Asymmetric Property

A ParentOf $B \rightarrow B$ ChildOf A : InverseOf Property

$(A \text{ ChildOf } B) \wedge (B \text{ ParentOf } C) \rightarrow (A \text{ Sibling } C)$: Chain Axiom Property

Consider a scenario where a newborn, Frank, is introduced into a family. We assume a strong adversary who possesses knowledge of mitochondrial inheritance patterns, ontology constraints, and the existing family structure, yet remains uncertain about Frank's parentage. Additionally, medical personnel are interested in querying the maximum number of siblings within this family for medical statistical purposes. This query is among a restricted set of queries permitted on the graph database. This task falls to a data analyst whose trustworthiness is in question.

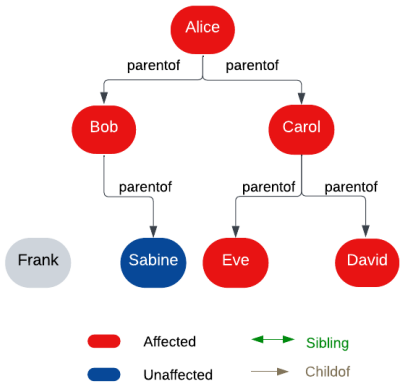

Figure 1: Family Graph Motivational Example

```
Q = SELECT (MAX(?siblingNb) AS ?maxSiblingCount)
WHERE {
    SELECT ?individual (COUNT(?sibling) AS ?siblingNb)
    WHERE {
      ?individual a <#Individual>.
      ?individual doc:sibling ?sibling.
    }
    GROUP BY ?individual  }
```

One adversary, unaware of the semantic axioms governing the graph, cannot draw conclusions regarding Frank's parentage based solely on this query result. Assuming a uniform distribution across all remaining individuals in the graph :

$\forall$ Q Posterior belief = Prior belief = 1/(NumberOfIndividuls-1)

Conversely, another adversary who is aware of the semantic axioms can draw probabilistic inferences from this query results:

Q=2 $\rightarrow$ Frank's parent can be: Bob, Eve, Sabine, or David.

Attacker conclusion: Frank has 25% chance of having mitochondrial disease.

Q=3 $\rightarrow$ Frank's parent can be: Alice or Carol.

Attacker conclusion: Frank has mitochondrial disease with 100% certainty.

One way to protect individual relationships is to answer the query in a manner that preserves privacy against a semantic-aware attacker.

3 DIFFERENTIAL PRIVACY

Differential privacy [1] provides a robust solution for such situations by ensuring that the inclusion or exclusion of any individual's data does not significantly affect the outcome of a query or analysis.

3.1 Differential Privacy Definition

A mechanism (Algorithm, System, Process..) M satisfies ε-differential privacy if for all neighboring data sets D_1 and D_2, and for all subsets S of the range of M, we have:

$$\Pr[M(D_1) \in S] \leq e^{\varepsilon} \times \Pr[M(D_2) \in S]$$

One way to achieve differential privacy (DP) is by computing the query sensitivity[4], which is defined as the maximum difference between all query responses of neighboring databases of the original one. Based on this sensitivity, noise is then added to query result, calibrated according to both the query sensitivity and the privacy budget, denoted by epsilon (ϵ). Epsilon serves as a balance between privacy and utility. Therefore, the selection of epsilon is a critical decision [3].

3.2 Differential Privacy on Knowledge Graphs

When applying differential privacy (DP) to relational databases, two databases are considered neighboring if they differ by only one record. However, when dealing with graphs that consist of nodes and edges, the concept of neighboring databases needs to be redefined. Many researchers have proposed effective notions of neighboring distances for graphs [2], including:

- **K-Edge-DP:** Defines neighboring graphs as those differing by k edges, with Edge-DP being a special case where k equals one.
- **Node-DP:** Two graphs are considered neighboring if they differ by one arbitrary node and all its incident edges.

These notions have proven effective in protecting graph data. However, the following work will show that they present privacy breaches when dealing with graph that follows OWL constraints.

4 LIMITATION OF DIFFERENTIAL PRIVACY IN SEMANTIC GRAPHS

When generating all graph neighbors with edge-dp distance, we initially obtain 49 neighbors for the graph depicted in figure 1. After filtering out neighbors that do not comply with OWL constraints and those that are illogical (e.g., self-parenting), we are left with 37 valid neighbors. Figure 3 illustrates examples of both valid and invalid neighbors (the invalid graph results from the OWL asymmetric property).

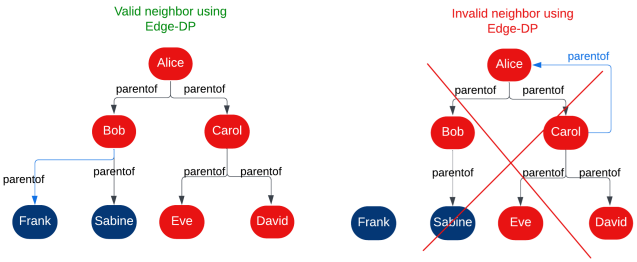

Figure 2: Edge-dp neighboring graphs

Consequently, the number of graphs that the attacker could confuse with the original has decreased. This reduction in neighborhood size can impact the effectiveness of protection, as a smaller set of neighboring graphs may make it easier for an attacker to distinguish the original graph.

We have also detected another issue where an attacker may be able to infer the true structure of the graph without even submitting a query. For instance, let's suppose now that Frank is indeed the child of Bob and that the graph follow an additional rule indicating the minimum number of children in this graph is two. This knowledge constraints the graph structure, requiring that each parent node, such as Bob, must have at least two children.

OWLDescription: ObjectProperty(parentOf, minCardinality(2))

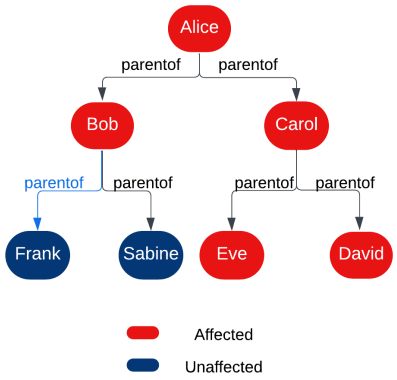

Figure 3: The only valid neighbor

By incorporating this knowledge, all neighbors are eliminated except for the one depicted in Figure 3, Frank's privacy is completely compromised, and the adversary's prior belief about his status (unaffected) becomes 100% accurate, even without the need to apply any queries.

5 CONCLUSION

For RDF graphs that follow OWL constraints, we have demonstrated that existing privacy concepts for protecting graphs are insufficient. In this PhD research, we aim to utilize graph data with ontology semantics to establish an optimal neighbor distance for ensuring differential privacy. Additionally, we will develop methodologies and metrics to quantify the loss of data privacy due to the incorporation of ontologies into graphs.

6 ACKNOWLEDGMENTS

Ph.D. Work funded by CyberINSA (France 2030 ANR-23-CMAS-0019)

REFERENCES

[1] Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*. Springer, 1–12.
[2] Michael Hay, Chao Li, Gerome Miklau, and David Jensen. 2009. Accurate estimation of the degree distribution of private networks. In *2009 Ninth IEEE International Conference on Data Mining*. IEEE, 169–178.
[3] Jaewoo Lee and Chris Clifton. 2011. How much is enough? choosing ε for differential privacy. In *Information Security: 14th International Conference, ISC 2011, Xi'an, China, October 26-29, 2011. Proceedings 14*. Springer, 325–340.
[4] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2007. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. 75–84.