Knowledge Graph Anonymization using Semantic Anatomization

Maxime Thouvenot[1,2],Olivier Curé[1], Philippe Calvez[2]

[1] Université Paris Est, LIGM, France.
{firstname.lastname}@u-pem.fr
[2] ENGIE CRIGEN - LAB CSAI
philippe.calvez1@engie.com

Abstract. Organizations and companies that are using Knowledge Graphs need to consider the privacy preservation of the data they contain. This is generally performed by anonymization techniques such as triple suppression and generalization which are known to reduce the utility of the released datasets. This paper presents semantic anatomization, a novel anonymization technique, that retains all quasi-identifier and sensitive values in the RDF graph. Due to an aggregating mechanism and the exploitation of the semantics contained in ontologies, this technique preserves data correlation and supports high quality analysis from anonymized graphs. We demonstrate the potential of semantic anatomization in a real-world setting on client information of a multinational company.

1 Introduction

As the usage of Knowledge Graphs (KGs) is going mainstream, it becomes necessary for organizations and companies to consider privacy-preserving data publishing (PPDP) approaches. In PPDP, we refer to the targets of an anonymization technique as "entities of interest" (*EoI*). Each EoI is identified by a set of values called attributes which can be split into four distinct categories: (i) Explicit identifiers (*EID*) explicitly identify an entity (*e.g.*, a social security number), (ii) Quasi identifiers (*QID*) are sets of attributes that together can potentially identify an entity, *e.g.*, date of birth, zipcode and gender, (iii) Sensitive attributes (*SA*) describe some sensitive information concerning an entity, *e.g.*, disease, political opinion, and (iv) Non-sensitive attributes (*NSA*) do not belong to any of the previous categories. EID, QID and SA can be considered as private attributes of the EoI while NSA are generally not concerned with privacy issues.

In the context of data represented using the RDF data model, several approaches and systems have been proposed. The anonymisation techniques that they are using mainly concern the suppression of information [1] [2], *i.e.*, removal of some triples or substitution of some RDF terms by blank nodes, and the generalization of some values [3]. In general, the adoption of these techniques favor the non-disclosure of entities over the utility aspect, *i.e.*, usefulness of the anonymized dataset to a group of users.

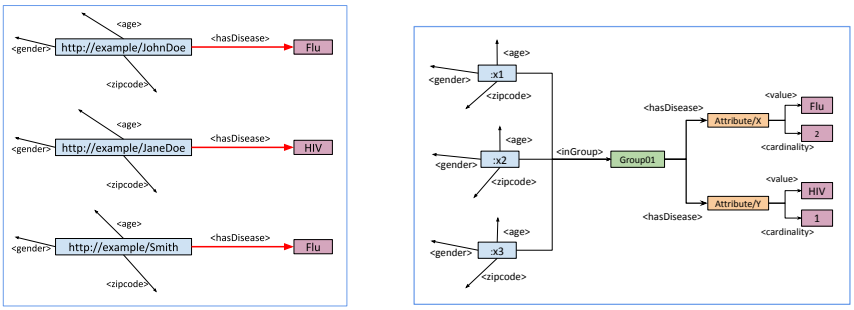
In this work, we have developed an anonymisation technique that focuses on the utility aspect but still provides high privacy guarantees. Denoted semantic anatomization, rather than suppressing or generalizing entity QIDs, this technique focuses on a semantic-aware grouping of SA. Additionally, it quantifies the amount of individuals in each group. As a result, it preserves the correlation between precise entity QIDs and semantically related SAs.

2 Semantic anatomization

Semantic anatomization adapts to KGs the main idea of the anatomy anonymization technique which has been applied to relational database [6]. To the best of our knowledge, this is the first implementation of an anatomy-based approach for KGs (note that in an RDF graph context, [5] only mentions it as a possible solution).

Adapted to RDF graphs, anatomy implies that the relationships between the QIDs and their SAs are removed by inserting intermediate nodes which will be used to gather different SAs into groups. Unlike other anonymization methods, anatomization only modifies the graph at the structural level by adding new nodes and edges. The fact that QID values are not transformed allows us to avoid the information loss that would be caused by some generalization mechanism and thus favor a quality analysis of the released data.

Fig. 1: Anatomization: removal of direct links between EoI and SA



(a) Original graph extract

(b) Application of the approach

We now provide a comprehensive example. Fig. 1(a) presents an extract of an original graph representing three patients (John Doe, Jane Doe and Smith) and some of their QIDs. In Fig. 1(b), we present the anatomized graph of this extract. We can see that the QID values have not been modified, the EIDs (John, Jane and Smith's URI) have been replaced by blank nodes and the $hasDisease$ property does not point to a disease concept but to a sub-graph group which is used to break sensitive relationships. Several properties have been added to the original graph, namely $inGroup$, $value$ and

$cardinality$, which belong to our Anatomy ontology. The latter two keep information about SAs, their value and cardinality, *i.e.*, number of occurrences in the dataset.

Consider an adversary who has access to the data in Fig. 1(b), possesses personal details about Jane Doe (*e.g.*, her age, gender and zipcode) and knows that she appears in this dataset. The attacker knows that Jane belongs to *Group 1* but cannot infer any additional information about her disease. However, using the cardinalities, he is able to deduce that $\frac{2}{3}$ of the entities contained in the group suffer from the Flu (resp. $\frac{1}{3}$ from *HIV*). Hence, he can only make a probabilistic guess about the actual disease that Jane has. Nevertheless, considering the utility aspect of this anonymization approach, it is still possible to extract valuable statistics on the disease of a group of people with a given age, zipcode or gender.

The objective of a semantic-enabled anatomization approach is to make sure that each group contains SA entries that are semantically related, hence, we are using an ontology associated to the dataset. In particular, we are making an intensive usage of the TBox's concept hierarchy. By leveraging these structures and a similarity measure denoted *taxonomy similarity* [4], we are able to compare the different concepts in our dataset and gather similar concepts together.

The semantic anatomization algorithm we are presenting in this section computes groups of SAs. Its result is used to create SPARQL update queries.

We chose to adopt a bottom-up approach which can be divided into two parts. First, we retrieve a set of initial clusters *i.e.*, clusters comprised of only one SA. Second, a merging step is performed and consists of the following operations: storing of all the clusters in a list L then we compute the *taxonomy similarity* between all these clusters. Similar clusters are merged and are thus defined by a set of concepts over which we search for a least common ancestor (LCA). This LCA is then used as the concept for the newly formed cluster. The process stops once every cluster contains at least 2 attributes. Note that the algorithm mainly takes as input the ontology and not the original dataset (except for retrieving initial clusters). This is important when considering the processing complexity of an algorithm since TBoxes are known to be orders of magnitude smaller than ABoxes (this is particularly true in our anonymization context). In fact, the complexity of our algorithm is $O(n^2)$ where n is the number of clusters computed from the ontology.

Once the final clusters are computed, we can produce the update operations to be performed to apply anatomization. We iterate on each of the clusters and on each of the attributes they contain in order to create the appropriate SPARQL queries (query template as well as a detailed clustering algorithm are available on GitHub[3]). Considering the graph of Fig 1(a), Its execution produces the graph presented in Fig. 1(b). Hence a deterministic set of queries is executed to produce the released dataset.

3 Evaluation

We now provide a preliminary evaluation of our approach on its effectiveness to fulfill the end-users' analysis need, *e.g.*, counting queries. The experimentation is conducted

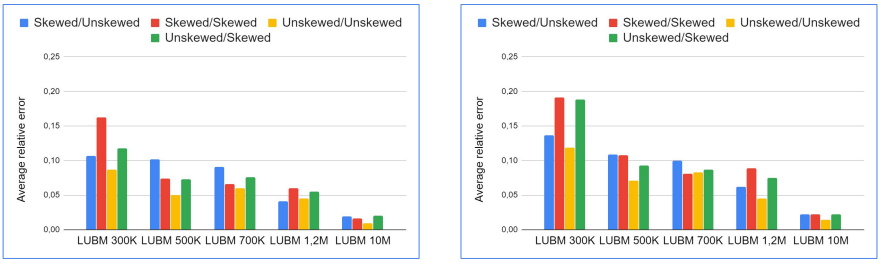
[3] https://github.com/mthouv/SemanticAnatomization

over an extension of LUBM datasets. In order to support a concept-based SA, we introduced a four-level concept hierarchy related to religion and assigned a religion to each person instance in our different datasets (ranging from 300,000 to 1,000,000 triples). We also added another attribute with no underlying ontology. Finally, we experimented with different value distributions for theses attributes which are displayed in the GitHub companion of this paper.

In order to measure the efficiency of the approach, we designed a set of counting queries involving 2 to 4 attributes (including 1 or 2 SAs) and computed the *average relative error* in answering these queries. The relative error can be described as $\mid Act - Est \mid /Act$ where *Act* and *Est* correspond respectively to the actual value derived from the dataset and an estimated value computed from the anonymized graph.

We provide the results for queries with 2 attributes in Figure 2 and show that the average error is below 10%. This is much better than what is generally achieved with generalization or suppression. Due to space constraint, the rest of the results can be found on our GitHub repository.

Fig. 2: Evaluation of the relative error



(a) Two attributes including one SA

(b) Two attributes including two SAs

Regardless of the various parameters (queries' size, value distribution, etc...), a common fact is observed: the larger the dataset (and consequently the more EoI we have), the more precise our approach is. This is due to the lesser selectivity of queries on larger datasets, *i.e.*, we compute our estimates on more significant subsets of the EoI. Most importantly, we showed that, even in an anatomized KG, it is still possible to retrieve valuable information. Furthermore, anatomization is not greatly affected by the distribution of SA values and becomes more efficient as the number of EoI grows. We observe slightly worse results when dealing with queries involving multiple SA but this is expected considering the method used to compute our approximation. Finally, query cardinality is not a limitation either.

4 Conclusion

We presented semantic anatomization as a new anonymization technique to help in the PPDP process for KGs. This new technique reaches a data analysis quality that

can hardly be met with the standard generalization and suppression techniques. This is mainly due to the preservation of data correlation between QIDs and SAs. Nevertheless, generalization provides privacy guarantees that would be hard to obtain with any form of anatomization. Hence, a future work is to integrate generalization in our system.

References

1. B. Cuenca Grau and E. V. Kostylev. Logical foundations of privacy-preserving publishing of linked data. In D. Schuurmans and M. P. Wellman, editors, *AAAI*, pages 943–949, 2016.
2. R. Delanaux, A. Bonifati, M. Rousset, and R. Thion. Query-based linked data anonymization. In *ISWC 2018*, pages 530–546, 2018.
3. B. Heitmann, F. Hermsen, and S. Decker. k-RDF-neighbourhood anonymity: Combining structural and attribute-based anonymisation for linked data. In *PrivOn at ISWC*, 2017.
4. A. Maedche and V. Zacharias. Clustering ontology-based metadata in the semantic web. In *PKDD 2002*, pages 348–360, 2002.
5. F. Radulovic, R. García-Castro, and A. Gómez-Pérez. Towards the anonymisation of RDF data. In *SEKE*, pages 646–651, 2015.
6. X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In *Proceedings of VLDB, Seoul, Korea, September 12-15, 2006*, pages 139–150, 2006.