Towards the Use of Graph Summaries for Privacy Enhancing Release and Querying of Linked Data

Benjamin Heitmann, Felix Hermsen, and Stefan Decker

Informatik 5 – Information Systems
RWTH Aachen University, Ahornstr. 55, 52056 Aachen, Germany
lastname@dbis.rwth-aachen.de

Abstract. Linked Data has become an important standard to describe meta-data about open government data. At the same time, most government data is not released as Linked Data. One reason for this could be the difficulty of applying privacy enhancing technologies such as differential privacy and private information retrieval to Linked Data. We introduce the idea of graph summaries to function as a schema for Linked Data which is schema-less. This in turn can provide a conceptual bridge for applying differential privacy and private information retrieval to Linked Data.

1 Introduction

Linked Data using the DCAT vocabulary is emerging as the de-facto standard for publishing meta-data about data sets released by governments, especially in the EU. At the same time, most of the data sets described by the DCAT meta-data are not being published using Linked Data.

This might be due to the fact that there currently is a lack of best practices for publishing sensitive data sets as Linked Data. For tabular data, established methods in the area of differential privacy [1, 2] exist, which preserves the utility of the data set while limiting the recoverability of personally identifiable information and avoiding the issue of de-anonymising the data set.

While Aron [3] makes a suggestion for an approach to apply differential privacy to Linked Data, one of the main short-comings is that the approach requires an *a priori* list of classes or properties to protect.

This is due to one of the main obstacles for applying the idea of differential privacy to Linked Data: Linked Data is schema-less, meaning that one data set can contain properties and classes from many vocabularies and ontologies. In addition, *a priori* knowledge about the structure of a linked data set is usually not possible. Due to this flexible and dynamic nature of LD, applying differential privacy like in [3] requires inspecting a data source and listing the sensitive classes and properties.

In this paper, we will introduce graph summaries as a central enabler for privacy enhancing release and querying of Linked Data. We argue that future best practices for identifying and masking the sensitive parts of an LD data set will incorporate some form of graph summarisation. In addition, we will explain why graph summaries could be very useful in enabling private information retrieval by distributing a query which can be fulfilled by one data source to multiple data sources.

The remainder of this paper is structured as follows: In section 2 we will introduce the background in regards to differential privacy and private information retrieval. Then in section 3 we will describe an approach for graph summaries which was introduced by Campinas et al. in [4]. Then in section 4, we describe how such graph summaries could be used to aid with the anonymised release of Linked Data in accordance with Aron [3]. In section 5 we describe how private information retrieval from federated SPARQL endpoints is enabled by graph summaries. We then conclude the paper and discuss future work in section 7.

2 Background

The European Data Portal[1] is publishing descriptions of data sets using the DCAT vocabulary. In particular, the DCAT-AP profile is used on the site. In terms of size, currently descriptions of approx. 600,000 data sets from around 70 data portals of 34 countries are hosted by the European Data Portal. Yet, only 1500 of these data sets are declared as explicitly using a form of RDF.

This suggests a mismatch between the requirements for releasing open government data and the tools and best practices available for publishing Linked Data. In particular, this could point to the difficulty of anonymising Linked Data sets before releasing them.

The release of anonymised, statistical data sets in tabular form has been solved to the most part using approaches which can be subsumed under the heading of differential privacy, which we describe below.

This is followed by a description of private information retrieval (PIR). PIR becomes relevant in relation to Linked Data, as SPARQL queries can reveal a lot about the intentions and interests of the query issuing party.

2.1 Differential privacy

The goal of releasing an anonymised statistical data, is to release data of a representative sample of a population in order to enable analysis of the common properties of that population, while keeping the properties of individual records secret.

The two most common elements of differential privacy approaches are generalisation and suppression in order to reduce the specificity of the quasi-identifiers [1]. In addition, a traditional and simple approach has been the randomization method [1] which can be explained as the modification of data at collection time by adding some noise to the records. If the noise is uniformly distributed,

[1] http://www.europeandataportal.eu/

the randomization method benefits from the fact that general data set properties (such as mean) stay invariant, while the perturbation of the records makes an identification hard.

Nevertheless there is a trade off between the utility of the data and the level of anonymization. On one hand, a great perturbation range gives a good randomization but also alters the records at a great extend. Clearly, the data become less expressive. On the other hand, a low bias retains the relevant information, but outlying records mostly remain unvaried and an attacker can identify these easily.

Aron [3] describes an approach for differential privacy for RDF data, by adding noise to Linked Data without changing the statistical properties of the data. However, he points out that two important issues are remaining. These are the issues of how to identify the representation of an individual record in the data set, and the issue of identifying which properties are of a sensitive nature and need to be protected.

As we describe in Section 4, graph summaries can provide a source of *a priori* knowledge about a linked data set, and have the potential to automate the application of differential privacy to Linked Data.

2.2 Private information retrieval (PIR)

The main idea of Private Information Retrieval (PIR) is that the user requests a particular element of the database without the database owner knowing which element the user was interested in [5]. There is a distinction between single- and multi-server PIR. In the single-server setup, the information is stored in only one place whereas, in the multi-server setup, multiple databases contain the same information thus allowing the user to load different parts of the data from various servers.

The main obstacle for traditional (non-SPARQL) multi-server PIR schemes, is that all instances of the database would need to contain the same information. This traditionally did imply the same organisation is in control of all the database instances with its data.

However, this scenario makes collusion of all database instances trivial. Because of this, more focus was put on research of the single-server PIR scenario. One of the results was an approach [6] which is fast enough to stream binary files and can saturate a 100Mbits/s line using a contemporary laptop.

In contrast, for federated SPARQL queries the assumption is that each SPARQL endpoint is controlled by a different organisation. This would make collusion much more unlikely. However, on the other hand, the data which is stored in a SPARQL endpoint is much more dynamic [7].

PIR becomes relevant in relation to Linked Data, as SPARQL queries can reveal a lot about the intentions and interests of the query issuing party. If we imagine a hypothetical use case in which open government data is hosted only on the server of a government agency, and made only available via a SPARQL endpoint, then that government agency can monitor all queries to the SPARQL endpoint. If the data were available “in bulk”, e.g. as an NTriples file, then all the privacy concerns can be trivially addressed by just downloading the data file and performing all queries locally. However, a malicious data provider will make sure that the data is not available for download or missing important parts.

In such a use case, the government could fully monitor all queries, e.g. made by an NGO or a journalist, as only the SPARQL endpoint provided by the government agency contains the data required to answer the query.

As we discuss in Section 5, the availability of graph summaries for SPARQL endpoints has the potential to enable splitting up a query which can be answered by one SPARQL endpoint to multiple SPARQL endpoints. In our use case, this would enable an NGO or a journalist to hide the details of their query and this their intent from the organisations hosting the SPARQL endpoints.

3 Graph summaries

As explained in section 2, in order to apply privacy enhancing approaches to Linked Data, such as differential privacy or private information retrieval, the contents of the Linked Data sets needs to be known in advance. This is due to the fact that Linked Data is schema-less and fast changing.

Graph summaries can provide a description of the graph, which contains the structure, i.e. the types of links and the classes of entities. As such, it can take the role of a schema as it is used for instance in relational databases. In addition, a graph summary has the benefit of not containing any of the actual information from the graph it summarises, if we assume that the class names and link names contain no information.

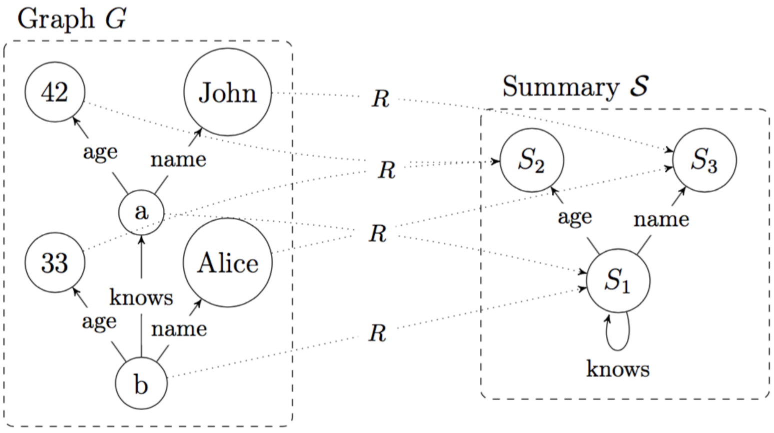

Fig. 1. Example of a graph and one possible summary from [8]. Dotted lines labelled R represent the summarisation relation that maps nodes in the graph G to nodes in S.

Campinas et al. [4] introduce a framework for graph summaries. The framework provides different approaches to summarise a graph, which focus on either types, properties or the structure. Figure 1 shows an example of a graph summary focusing on the properties.

Campinas et al. [4] present algorithms for both precise and approximate graph summaries. Both types of summaries provide a graph which is homomorphic to the original graph. However, while *precise graph summaries* contain every graph which is also contained in the original graph, *approximate graph summaries* are more robust in the face of graphs with errors and inconsistencies.

In addition, for precise graph summaries the worst case for the size is to be almost as big as the original graph. In contrast, approximate graph summaries are much smaller.

The steps of the algorithm are as follows, more details are in [8].

Gathering of entity descriptions: The description of all entities is collected with one complete pass over all edges of the graph. This provides the necessary contextual information needed for the next step.

Mapping of all nodes to their respective nodes in the summary graph For every node, a corresponding summary node is either created or selected from the list of already generated summary nodes.

Materialisation of summary edges Another pass over all edges of the original graph, in order to decide if that edge needs to be materialised in the summary graph.

After providing formal definitions for both types of summaries, two implementations for calculating graph summaries are given in [4]. The first version uses SPARQL queries, while the second version can be run on shared-nothing computation platforms such as Hadoop.

Campinas provides the results of an evaluation on over 14 real-world datasets of various size and complexity in Section 5.3 of [8]. The results of the evaluation show that Campinas algorithms make it possible to approximate the errorless summary graph quite accurately but with a much lower space and time complexity. In addition, the evaluation supports the claim that the algorithm can be used on any kind of Linked Data.

In comparison to other existing graph summarisation approaches, the approach presented by Campinas in [8] and [4] is flexible and expressive using graph homomorphisms, whereas other approaches are designing for specific applications. No details of the graph need to be known beforehand, as the algorithm is solely based on the features of the entities. The algorithm requires multiple iterations, so the performance is dependant on the size of the graph.

In summary, the approach for graph summaries proposed by Campinas et al. can provide a schema for the schema-less Web of Data, while also having the benefit of not revealing any other properties of a data set.

4 Using graph summaries for the release of sensitive Linked Data

Graph summaries can provide a source of *a priori* knowledge about a linked data set, and have the potential to automate the application of differential privacy to Linked Data.

As described by Aron [3], in order to apply differential privacy to Linked Data, it is necessary to know how to identify the enitites which need to be protected. The availability of a graph summary allows heuristics to be now used to identify such entities, based on existing definitions of k-anonymity, l-diversity and t-closeness [9] [10], which we can paraphrase as follows for the graph summary:

k-anonymity A graph satisfies k-anonymity, if for every instance of a class C, there are k-1 other instances, so that their quasi-identifiable attributes have equal values with the other instances of class C.

l-diversity A graph satisfies l-diversity, if for every class C, there are at least l values for every property of every instance of class C.

t-closeness A graph satisfies t-closeness if the distance between the distribution of every property of every class C in the original graph and the anonymised graph is not bigger than a threshold t.

While the use of graph summaries has potential to allow the whole process of anonymising and masking a linked data graph to be automated, there are many reasons to still involve human inspection of the final data set. In particular, to prove that the responsibilities of an organisation have been fulfilled.

5 Using graph summaries for private information retrieval

The availability of graph summaries for SPARQL endpoints has the potential to enable splitting up of queries which can be answered by one SPARQL endpoint to multiple SPARQL endpoints. This would enable an NGO or a journalist to hide the details of their query and this their intent from the organisations hosting the SPARQL endpoints.

In particular, with the availability of graph summaries for all available SPARQL endpoints, we can imagine the role of the *query planner* to be almost the opposite of query planners which try to optimise the query response time. For private information retrieval, the query planner will attempt to distribute a query across as many SPARQL endpoints as possible. The intended goal then is to make it impossible for any observer with only partial knowledge of the query to determine the goal of the query. This mode of operation could also be called an *obfuscating query planner*.

6 Analysis of threat model

Deng et al. [11] introduced a framework for analysing threat models. Using their classification of threats, our suggestions for using graph summaries have to be classified as “hard privacy” approaches, as they are limiting the release of data.

First we list how the use of graph summaries for differential privacy addresses the hard privacy threats as defined by Deng et al. in [11]:

Linkability: Using graph summaries allows removing of personally identifiable data, which mitigates the threat of linking an entity in the anonymised graph with an entity in the original RDF graph.

Identifiability: As entities in the original and anonymised graph are not linkable, the identity of persons is also protected.

Non-repudiation: In addition, as identities are protected, there should be no proof of a persons information being part of the anonymised data set.

Detectability: In a similar way, participation of a person in the anonymised data set should be undetectable.

Disclosure of information: Finally, as this approach removes sensitive data before it is released, disclosure of sensitive information is not possible.

In summary, using graph summaries for differential privacy maintains unlinkability and anonymity of identities in the original data set, and it limits unintended disclosure of information. However, it does not address non-repudiation and detectability.

Next we list how the use of graph summaries for private information retrieval address hard privacy threats:

Linkability: Using an obfuscating query planner makes the different parts of the query unlinkable to each other, provided there is no collusion between the different SPARQL endpoints.

Identifiability: Suitable credentials will be required in order to access all required SPARQL endpoints, if no anonymous access without credentials is available. Therefore the person or organisations initiating the queries is always known.

Non-repudiation: As standard SPARQL access is used, it is not possible to deny having initiated the SPARQL queries.

Detectability: In addition, as SPARQL queries are not sent over an encrypted channel, any potential eavesdropper can listen to the SPARQL queries.

Disclosure of information: The information which is hidden from SPARQL endpoints is the intent of the initiator and the original SPARQL query. However eavesdroppers of the SPARQL queries could try to re-assemble the original SPARQL query.

In summary, using graph summaries for private information retrieval, hides the intent behind the split SPARQL queries as the queries are unlinkable if the SPARQL endpoint operators do not collude. In addition, the original query and its intend are not disclosed. However, the privacy threats of identifiability, non-repudiation and detectability are not addressed.

7 Conclusion and future work

We have introduced the idea of graph summaries as a privacy enhancing technology to enable differential privacy and private information retrieval.

As graph summaries contain no information about a graph beyond the classes and properties of the original data set, there is no leakage of personally identifiable information. Graph summaries can be used almost like the schema for traditional, relational databases. As such they have the potential to provide a conceptual bridge for applying many existing privacy enhancing technologies from relational databases to Linked Data.

In terms of future work, we are planning to implement the graph summary approach as part of a framework for automatic anonymisation of Linked Data sets. In particular, we will develop a testbed for evaluating different parameterisations of anonymised release of Linked Data, against several state of the art algorithms for deanonymising graph data as described in the overview of Al Azizy et al. [12]. In addition, we are planning to implement a query processor for obfuscated query planning to enable private information retrieval using graph summaries.

References

1. Venkatasubramanian, S.: Measures of anonymity. In: Privacy-Preserving Data Mining. Springer (2008) 81–103
2. Dwork, C.: Differential privacy: A survey of results. In: International Conference on Theory and Applications of Models of Computation, Springer (2008) 1–19
3. Aron, Y.: Information privacy for linked data. Master's thesis, MIT CSAIL (2012)
4. Campinas, S., Delbru, R., Tummarello, G.: Efficiency and precision trade-offs in graph summary algorithms. In: Proceedings of the 17th International Database Engineering & Applications Symposium, ACM (2013) 38–47
5. Yekhanin, S.: Private information retrieval. Communications of the ACM **53**(4) (2010) 68–73
6. Aguilar-Melchor, C., Barrier, J., Fousse, L., Killijian, M.O.: Xpir: Private information retrieval for everyone. Proceedings on Privacy Enhancing Technologies **2016**(2) (2015) 155–174
7. Buil-Aranda, C., Hogan, A., Umbrich, J., Vandenbussche, P.Y.: Sparql web-querying infrastructure: Ready for action? In: International Semantic Web Conference, Springer (2013) 277–293
8. Campinas, S.: Making Sense of Web Data. PhD thesis, National University of Ireland, Galway (2016 (to appear))
9. Ciriani, V., di Vimercati, S.D.C., Foresti, S., Samarati, P.: κ-anonymity. In: Secure data management in decentralized systems. Springer (2007) 323–353
10. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: 2007 IEEE 23rd International Conference on Data Engineering, IEEE (2007) 106–115
11. Deng, M., Wuyts, K., Scandariato, R., Preneel, B., Joosen, W.: A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements. Requirements Engineering **16**(1) (2011) 3–32

12. Al-Azizy, D., Millard, D., Symeonidis, I., O'Hara, K., Shadbolt, N.: A literature survey and classifications on data deanonymisation. In: International Conference on Risks and Security of Internet and Systems, Springer (2015) 36–51