# Semantic Web Enabled Record Linkage Attacks on Anonymized Data

Jacob Miracle and Michelle Cheatham

DaSe Lab, Wright State University, Dayton, OH 45435, USA,
miracle.13@wright.edu,
michelle.cheatham@wright.edu

**Abstract.** Big Data analytics holds the promise of enabling new discoveries in medicine, more efficient business practices, and other important advances. However, much of the data involved in such analyses contains personally identifiable information (PII) that needs to be removed or obscured prior to release in order to protect individuals' privacy. Anonymizing a dataset is not as easy as it seems, however, and many supposedly anonymous datasets are vulnerable to a record linkage attack. Most of these attacks are currently conducted manually and can be labor-intensive, but as semantic web technologies continue to gain popularity, the potential for automating various aspects of these attacks increases. This paper explores the components of a record linkage attack and how semantic web technologies could play a role in facilitating them.

**Keywords:** Anonymization, Semantic Web, Privacy, De-anonymization Attacks

## 1 Introduction

Humanity is producing more data than ever before, and in the hands of researchers this data has led to important insights in a variety of domains, from medicine to construction to marketing. Privacy concerns are obviously an issue in this environment, and most data sets that are made public or used in research have been through some sort of anonymization process. Unfortunately, this process frequently consists entirely of removing explicit identifiers, such as name, email address, and social security number, from the data. In many cases, such an anonymization process still leaves the data vulnerable to a record linkage attack.

Consider a dataset that contains the following fields: name, social security number, job title, gender, age, zip code, and salary. A company wishes to make this dataset available to the media to show that there is no pay gap between men and women in the organization. To avoid revealing the salary of individual employees, the names and social security numbers are removed. However, the remaining fields can still be used to link individuals to their salaries given the presence of an appropriate secondary dataset. For example, if the company also keeps its employees' CVs on its website that contain their names, job titles, and addresses (and a person's gender can often be inferred from their name), then

it might be possible to link the two data sources and thereby associate names with salaries. The privacy of employees who have an uncommon job title or live in a sparsely populated zip code is particularly at risk in this scenario.

Tables 2a and 2b show an example of a record linkage attack involving these datasets. In this case, gender, job title, and zip code are *quasi-identifiers* that can be used to link records across the two datasets. For example, it can be inferred that Jonathan Wilson makes $54,750 because he is the only male Software Developer living in the 24932 zip code, but it is not known who makes $37,500 because there are several people with the same combination of gender, job title, and zip code: Abby Johnson, Victoria Stevens, and Stephanie Lewis. The risk to a particular person's privacy depends on the number of people who share that individual's quasi-identifier values. Latanya Sweeney called this number of people k and established the concept of k-anonymity, in which the values in a dataset are generalized such that there are at least k records for each combination of quasi-identifier values [13]. In this example, if only the first four digits of the zip codes are included in the data, then Jonathan Wilson's salary can no longer be determined, because he is now indistinguishable from another person. This increased privacy comes at a cost – since the full zip codes are no longer included in the data, an analysis of a gender equality for salary can no longer control for differences in local standard of living costs as accurately.

**Table 1.** Record Linkage Attack

| Job Title | Gender | Age | Zip Code | Salary |
|---|---|---|---|---|
| Software Developer | Male | 32 | 24932 | $54,750 |
| Software Developer | Male | 32 | 24937 | $64,200 |
| Hardware Developer | Female | 30 | 24944 | $37,500 |
| Hardware Developer | Female | 30 | 24944 | $45,500 |
| Hardware Developer | Female | 30 | 24944 | $55,600 |

(a) Anonymous Dataset

| Name | Job Title | Gender | Age | Zip Code |
|---|---|---|---|---|
| Jonathan Wilson | Software Developer | Male | 32 | 24932 |
| James O'Brien | Software Developer | Male | 32 | 24937 |
| Abby Johnson | Hardware Developer | Female | 30 | 24944 |
| Victoria Stevens | Hardware Developer | Female | 30 | 24944 |
| Stephanie Lewis | Hardware Developer | Female | 30 | 24944 |

(b) Identified CV Dataset

Other researchers have since expanded upon the initial model of k-anonymity [1, 10, 9]. The underlying idea remains the same throughout these works: even when explicit identifiers have been removed from a dataset, some identities may be discovered if another dataset that contains explicit identifiers shares some fields and individuals with the anonymous dataset can be found. Finding such

a dataset is possible in a surprising number of circumstances. In our own work, we have deanonymized annual workforce surveys from the Ohio Board of Nursing by linking them with voter registration rolls and the state licensing website. Finding an appropriate dataset with which to link the target data can be difficult though, involving a lot of manual search and evaluation of possibilities. In this position paper we discuss how the widespread adoption of Semantic Web technologies may make conducting record linkage attacks quicker and easier in the future. Awareness of the potential negative as well as the positive uses of such technologies is an important first step towards designing and developing privacy safeguards on the Semantic Web.

## 2 Data Availability

Record linkage attacks require a dataset against which to link the anonymized data. Historically, most data was inconvenient to use since it was available only as databases or on file servers as spreadsheets, CSV files, or tables in PDF documents. Retrieving the data could also be difficult. For instance, some repositories might be accessible via websites or structured query mechanisms while others required a login and use of secure file transfer protocols. Financial drawbacks also inhibited data integration. Some data might be stored using proprietary formats that required expensive software licenses to read. These obstacles made finding and retrieving data related to an attacker's target dataset difficult.

The rise of linked data has changed this situation drastically. Linked data is expressed as RDF and can be accessed using standard protocols. It is also prolific. The Linked Open Data (LOD) Cloud now contains over 31 trillion triples across 295 datasets, with more than 503 million links across datasets [3]. The most represented domains in the cloud are social network information and government data, composing over 51% and 18%, respectively of the total [3]. This is a huge amount of information about many different aspects of people's lives, and the potential for its misuse should not go unconsidered. For example, Table 2 shows data from the Open University in the United Kingdom.[1] This information can be downloaded in various formats or accessed via a SPARQL endpoint. It includes information about a person's job title, groups they belong to, and publications they have co-authored. While this data is not generally considered sensitive, it could be used as a quasi-identifier for a target dataset. Additionally, some information included in this data, including usernames on social media platforms such as Twitter and LinkedIn and a list of other linked datasets in which this person appears, can be used to find more information about this person.

Another quickly growing type of data on the Semantic Web is that annotated with schema.org markup.[2] Schema.org is an initiative by major search engine companies to facilitate the description of entities and the relationships between them using a basic syntax expressed as RDFa, Microdata, or JSON-LD. As of 2014, more than 36% of websites in Google's crawl contained schema.org markup

[1] http://data.open.ac.uk/page/context/people/profiles
[2] https://schema.org

**Table 2.** A record from the Open University personal profile linked dataset.

| Property | Object |
|---|---|
| Title | Dr |
| Given name | James |
| Family name | Rees |
| Job title | Anthony Nutt Sr Research Fellow |
| inDataset | Open Research Online |
| inDataset | OU People Profiles |
| Account | https://www.linkedin.com/nhome/... |
| Holder of | http://data.open.ac.uk/role/ResearchStaff |
| Mailbox SHA1 sum | 4d1c4c2e8f5...34d55078a3f |
| Has membership | http://.../the-open-university-business-school |

[11]. It is particularly commonly used to describe people, businesses, events, and reviews. Fields relevant to people include many likely quasi-identifying fields, such as birth date, birth place, gender, nationality, and affiliation.

## 3 Data Relevance

While the rise of linked data and schema.org markup has made much more data available in an easily accessible manner, a record linkage attack relies on finding datasets that include relevant information about the individuals in the target dataset. Finding an appropriate dataset is often the most time-consuming aspect of a record linkage attack. However, some research currently underway can speed up this process, thereby lowering the barrier to deanonymization.

Many linked datasets have very complex or extremely simple schemas. It is often difficult for a potential user of a dataset to quickly identify whether or not the data is useful for their purpose, but numerous methods for summarizing a linked dataset speed up this process. For example, the linked data summarization approach described in [14] ranks the axioms within an ontology based on graph-based measures such as centrality, while Loupe is an online tool that provides statistics regarding the usage of properties to describe instances of particular types within a linked dataset.[3]

Visualization tools are another avenue for quickly determining the general content and structure of a dataset. Many visual interfaces for data exploration on the Semantic Web involve displaying the RDF data as a graph. Unfortunately, graph-based representations frequently place entities based on graph metrics such as centrality or density, rather than according to their semantic meaning. They also have difficulty scaling to large datasets without becoming unwieldy. Kow and his colleagues have attempted to move beyond this towards more semantic-based layout algorithms with their idea of an "information landscape," which places similar concepts near one another and labels clumps of

[3] http://loupe.linkeddata.es/loupe/

entities with terms describing the group. Users can select areas of interest that seem likely to contain relevant entities, which automatically filters the mappings shown in the list. This method of filtering allows users to systematically explore an ontology at a high level of detail without losing track of the big picture [8]. Other approaches handle the problem of scalability by providing an RDF triple browser interface rather than attempting to show the entire dataset at once [6].

Ontology Design Patterns (ODP) provide another means for quickly determining linked dataset relevance to attackers. ODPs are self-contained, reusable patterns that model concepts that commonly occur across different ontologies. A well designed ODP describes the key aspects, and only the key aspects, of the concept being modeled [4]. If an attacker can quickly isolate the part of a complex schema most related to an ODP of interest, or quickly determine whether or not data with little schema information fits into the ODP model, they would have a better idea of whether or not the dataset in question was useful. For example, a person's communications often reveal much about them. Blomqvist posted an ODP on the website ontologydesignpatterns.org to model a "Communications Event." A simplified illustration of this ODP is shown in Figure 1.

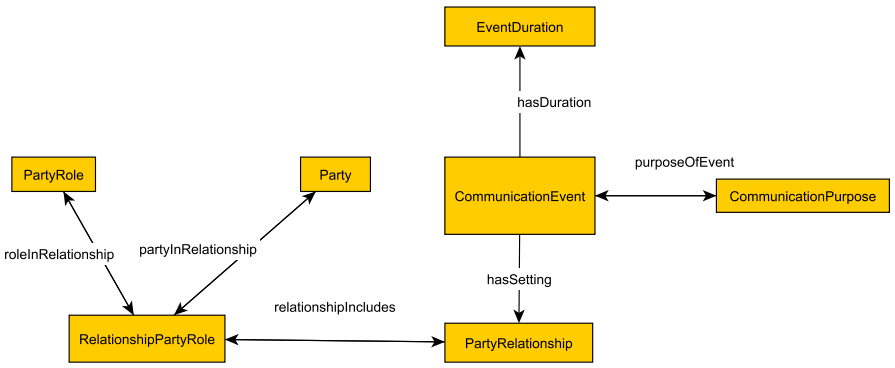


**Fig. 1.** A subset of the Communications Event ODP

The dataset containing information about the 2012 European Semantic Web Conference available at http://data.semanticweb.org/dumps/conferences/ contains information about, among other things, the keynote talks given at the conference, including their start and end times, the speaker, the topic, the setting in which the talk occurred, and the title and subject matter. A method for detecting the presence of an ODP (such as Communications Events) in a linked dataset was proposed by Khan and Blomqvist in [7]. For datasets with a significant schema, an ontology alignment system (described in Section 4) could also be used to recognize that this dataset contains Communications Events.

## 4 Data Linking

Once a relevant dataset has been identified, the target dataset must be joined with it based on the quasi-identifier values. This seems straightforward, but can actually be quite difficult in practice because the schemas for the two datasets were likely developed by different people, for different purposes. Because of this, even two ontologies that represent the same domain will generally not be the same. They may use synonyms for the same concept or the same word for different concepts, they may be at different levels of abstraction, they may not include all of the same concepts, and they may not even be in the same language. Furthermore, the classes and properties in the ontologies may not be used consistently when describing the entities within the dataset. The goal of ontology alignment is to determine when an entity in one ontology is semantically related to an entity in another ontology, despite these challenges.

The Ontology Alignment Evaluation Initiative (OAEI) is a set of benchmarks for evaluating the performance of alignment systems. The initiative has held evaluations annually since 2005. Over that time, the accuracy and the variety of problems handled by alignment systems have increased, while runtimes have decreased.[4] The top performing alignment systems include two that are available online: AgreementMakerLight[5] and LogMap.[6] These systems achieve an F-measure of .76 and .73, respectively, on an OAEI track based on aligning ontologies related to conference organization. These results are approaching the level of consensus that humans have when performing alignment tasks [5], implying that the dataset linking phase may be an aspect of record linkage attacks that could be automated in the near future. Additionally, alignment systems could be used to attempt to align datasets to ODPs representing key concepts, such as a Person, in order to refine a collection of possibly-relevant datasets for further analysis. Aligning a dataset against an ODP rather than another dataset can be easier, due to the limited scope and application-neutral nature of an ODP.

Coreference resolution algorithms attempt to determine when the same instance (i.e. individual) is referred to in two in different ways. For instance, is John Q Public in one dataset the same person as J.C. Publick in another? This is the Semantic Web technology most closely related to deanonymization: determining whether a person whose name and social security number have been replaced with random strings, for example, is present within an external dataset such as voter registration records is precisely what coreference resolution algorithms attempt. Most current approaches use string similarity metrics to compare two instances based on their property values (e.g. zip code, age, height) or their property values together with the names of those properties. Top performing systems on the mainbox instance matching task include the aforementioned LogMap and RiMOM [12], with F-measures of .83 and .91, respectively.

[4] http://oaei.ontologymatching.org
[5] https://github.com/AgreementMakerLight/AML-Jar
[6] http://csu6325.cs.ox.ac.uk

## 5 Data Inferencing

Traditional record linkage attacks sometimes involve specific or general knowledge or assumptions an attacker has about a target. This can be made significantly easier using automated reasoners. For example, assume the attacker is working with a medical records dataset organized according to a schema which includes the statements below.

```
<exSchema:hasDisease> <rdfs:domain> <exSchema:Person>
<exSchema:hasDisease> <rdfs:range> <exSchema:Disease>
<exSchema:LungDisease> <rdfs:subClassOf> <exSchema:Disease>
<exSchema:HeartDisease> <rdfs:subClassOf> <exSchema:Disease>
<exSchema:hasEthnicity> <rdfs:domain> <exSchema:Person>
<exSchema:hasEthnicity> <rdfs:range> <xsd:string>
<http://data.ex.org/person/12345> a <exSchema:Person>
<http://data.ex.org/person/12345> <rdfs:nameFull> "Zhang Lu"
```

If the attacker knows that Zhang Lu has some disease and wants to determine what it is, he can add some additional statements to the knowledge base that reflect his assumptions and then use a reasoner to check whether or not it is possible to infer the disease Mr. Zhang has, based on those assumptions. For instance, the attacker may know that Mr. Zhang is of Asian ancestry. He could then add the following fact to the knowledge base:

```
<http://data.ex.org/person/12345> <exSchema:hasEthnicity> "Asian"
```

The attacker could further assume that people of Asian ancestry are unlikely to get heart disease (based on statistical knowledge). The attacker would add the following fact to the knowledge base:

```
(exSchema:Person and (exSchema:hasEthnicity Asian)) SubClassOf:
not (exSchema:hasDisease some exSchema:HeartDisease)
```

The attacker could then use an automated reasoner to determine whether or not Mr. Zhang's disease could be inferred. While space constraints force this example to be relatively simplistic, it shows that the attacker can use existing Semantic Web languages and tools to quickly explore the ramifications of any assumptions he would like to make.

## 6 Conclusion

This position paper explored the potential for Semantic Web technologies to facilitate record linkage attacks against anonymized datasets. The emergence of linked data and schema.org markup has increased the availability of data that can be used to conduct attacks. Tools for data summarization and visualization can assist an attacker in sifting through this data to find a relevant dataset

with which to link the target dataset. Meanwhile, the emergence of automated techniques for ODP identification, ontology alignment, coreference resolution, and reasoners hold the potential to one day fully automate record linkage attacks.

The ramifications of Semantic Web technologies on privacy are likely to be profound. Aggarwal showed that typical approaches to anonymize data break down in the face of high dimensionality [1], which is precisely what linked data provides. Dealing with this may require difficult decisions about how to publish sensitive data, potentially involving perturbing the sensitive values [2], which has corresponding impacts on its utility. These concerns are present whether the sensitive data is published as linked data or in a database, CSV file, or other traditional format, because as we have seen, even innocuous data about a person available on the Semantic Web can be used to deanonymize a standalone dataset.

In future work on this topic, we plan to assess the volume of data containing potential quasi-identifiers that currently exists on the Semantic Web and develop a data vulnerability assessment tool based on Semantic Web technologies.

## References


1. Aggarwal, C.C.: On k-anonymity and the curse of dimensionality. In: Proceedings of the 31st International Conference on Very Large Databases. pp. 901–909 (2005)
2. Aggarwal, C.C., Philip, S.Y.: A general survey of privacy-preserving data mining models and algorithms. In: Privacy-preserving data mining, pp. 11–52. Springer (2008)
3. Anja Jentzsch, Richard Cyganiak, C.B.: State of the lod cloud (September 2011), http://lod-cloud.net/state/ [Online; accessed 29-February-2016]
4. Cheatham, M.: The Properties of Property Alignment on the Semantic Web. Ph.D. thesis, Wright State University (2014)
5. Cheatham, M., Hitzler, P.: Conference v2.0: An uncertain version of the oaei conference benchmark. International Semantic Web Conference pp. 33–48 (2014)
6. Erling, O., Mikhailov, I.: Faceted views over large-scale linked data. LDOW (2009)
7. Khan, M.T., Blomqvist, E.: Ontology design pattern detection-initial method and usage scenarios. In: SEMAPRO 2010, The Fourth International Conference on Advances in Semantic Processing. pp. 19–24 (2010)
8. Kow, W.O., Sabol, V., Granitzer, M., Kienrich, W., Lukose, D.: A visual soa-based ontology alignment tool. In: Proceedings of the 6th International Conference on Ontology Matching. pp. 242–243. CEUR-WS. org (2011)
9. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: ICDE (Purdue University) (2007)
10. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity (March 2007)
11. Mika, P., Potter, T.: Metadata statistics for a large web corpus. LDOW 937 (2012)
12. Shao, C., Hu, L.M., Li, J.Z., Wang, Z.C., Chung, T., Xia, J.B.: Rimom-im: A novel iterative framework for instance matching. Journal of Computer Science and Technology 31(1), 185–197 (2016)
13. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10 (May 2002)
14. Zhang, X., Cheng, G., Qu, Y.: Ontology summarization based on rdf sentence graph. In: Proceedings of the 16th International Conference on World Wide Web. pp. 707–716. ACM (2007)