Why Open Data May Threaten Your Privacy

Sebastian Pape
Goethe University Frankfurt
Chair of Mobile Business & Multilateral Security
Frankfurt, Germany
sebastian.pape@m-chair.de

Jetzabel Serna-Olvera
Goethe University Frankfurt
Chair of Mobile Business & Multilateral Security
Frankfurt, Germany
jetzabel.serna-olvera@m-chair.de

Welderufael B. Tesfay
Goethe University Frankfurt
Chair of Mobile Business & Multilateral Security
Frankfurt, Germany
welderufael.tesfay@m-chair.de

ABSTRACT

In this position paper we discuss the effect of open data on privacy. In order to reduce privacy issues due to the publication of open data, we suggest to build a database which overviews open data in a structured way with a special focus on privacy. This database could be enhanced with tools which automatically try to link existing datasets and allow publishers to check potential de-anonymization risks.

1. INTRODUCTION

Open data is the basis of a wide range of applications and services aimed at improving our daily lives. In particular, initial efforts of making data available by public administration bodies aimed at increasing accountability and transparency [13]. However, recently they are also providing these data in order to create new and more efficient services. Examples for those services are improved health services, mobility services based on open transportation data, and services with educational purposes [10].

In general, before this information is published, it is anonymized, e.g., personal identifiers are removed. The main goal of anonymization is that, analysts will still find the data useful while it is not possible to identify people whose information is included in the dataset. However, from a theoretical point of view, published data can only be either useful or perfectly anonymized [7, 17]. Indeed, since more than a decade, the de-anonymization of data is a known issue. The most famous examples are de-anonymization of health data published in Massachusetts [20], identifying Netflix users [15] or most recently identifying people by their credit card meta data [6]. The de-anonymization generally works as follows: sparse information in the dataset which is unique to the user (e.g., Zip code, sex and date of birth; the timestamp of a rating; or location and time of a purchase) is linked to external information with information on the person's id. We claim that probabilities of de-anonymization increase due to the availability of privacy related external data and not particularly because of the poor performance of certain anonymization algorithm. In this respect, this issue is further reinforced thanks to the raise of open data, open government initiatives and ultimately data publicly available on social networks. Adversaries have more and more information available, which, can be linked to re-identify people. Additionally, machine learning approaches amplify this, since they are able to work on fuzzy data (e.g., [12]). With machine learning approaches, data which is not directly linkable may be matched and probabilistic information on individuals may be derived. In practice, this is already done for credit scorings.

In order to detect and prevent privacy issues with open data, tools are needed to support researchers and publishers of open data [16]. Even though, there exist imminent risks derived from linking open data with data of different provenance, such as social networks or illegally leaked datasets due to security breaches; yet, we limit the scope of our research, to investigate methods and tools to support the process of publishing open data, by considering only open data initiatives. To this aim, we propose to build a database giving an overview of available open data in a structured way with a special focus on privacy. This database could be enhanced with mechanisms which automatically try to link existing datasets and detect potential privacy risks. This would allow publishers to check how their data could be linked before they actually publish the data.

To make some cases for our arguments, we give examples of publicly available data in Sect. 2 and provide a short survey on de-anonymization studies in Sect. 3.

2. OPEN DATA

There is tendency for public administrations to publish different sorts of open data, ranging from city administration datasets, to economy, urban environment, population and territory related data. We analyze the datasets published by three different entities: 1) the USA Central Intelligence Agency World Factbook; 2) the Barcelona Open Data Portal (OpenDataBCN); and, 3) the USA Government (USA.Gov). We also provide some insights which of them could potentially represent privacy risks. Risks arise either when data is aggregated with other available datasets or just by simple correlation with publicly available information, such as data provided by social networks.

PRINF '15 Dresden, Germany

The USA Central Intelligence Agency World Factbook [2] has made available a full collection of global data from 267 entities. For each entity, datasets include data associated to demographics, health, education, transportation, economy, society, and among others, politics. These datasets are well known because of its well defined structure and format. This allows semantic web researchers to apply a number of machine learning algorithms for semantic relation analysis, semantic search queries, unsupervised clustering, supervised learning and anomaly detection [22].

USA.gov [4] provides a dataset about the USA Gov web page accesses. Data is obtained every time anyone views any of its URLs. Attributes published in the dataset include: timestamp, location related data (i.e., country, city, timezone, and coordinates), language preferences, referring and clicked URL, and details about the user agent (e.g., browser and platforms).

Similar to the aforementioned initiatives, the BCN Open Data Portal [1] has made available 324 datasets. Among the most descriptive datasets, are those providing the demographics of the city of Barcelona. The demographic datasets contain information classified by district, gender, age range, nationality, academic level and even the number of people currently living alone. Although, in principle such data contains no personal identifiable information; combined datasets (e.g., number of foreign females by nationality and district, number of foreign females by age and district, and number of females by age and district who live alone), could be "easily" correlated with publicly available information. Moreover, advanced machine learning and de-anonymization techniques, could benefit from the combination of such datasets to obtain high probabilistic de-anonymization results.

3. DE-ANONYMIZATION STUDIES

In this section, we look into sample research efforts carried out in the subject of de-anonymization. Specifically, we surveyed what has been done methodologically and which datasets have been exploited to achieve the study results.

Narayanan and Shmatikov [15] presented a new methodology of de-anonymization attack against high dimensional micro-data. They provided a formal model for privacy breaches showcasing the fundamental limits of privacy in public micro-data. The authors also provided a practical analysis of the Netflix Prize dataset, containing anonymized movie ratings of 500,000 Netflix subscribers. By using the Internet Movie Database as the source of background knowledge to the Netflix dataset, they were able to successfully identify a number of specific members. Furthermore, the authors highlighted that de-anonymization requires data that is abundant, granular and fairly stable across time and context.

Sweeney [19] investigated the de-anonymization of US population data using attributes such as 5-digit ZIP, gender, date of birth, and place (i.e., city, town, or municipality). The author used two datasets for this study, namely, the voter registration list of Cambridge Massachusetts and medical record data, and linked the attributes to re-identify users in the list. In a similar study [21] she worked on the patient data of the state of Washington in the United States. This data contained hospitalizations that happened in the State in a given year, including patient demographics, diagnoses, procedures, attending physician, hospital, a summary of charges, and how the bill was paid. The dataset, however, did not contain patient names or addresses except for ZIPs. Newspaper stories printed in the same State for the same year that contained the word "hospitalized" often included a patient's name and residential information and explained why the person was hospitalized, such as vehicle accident or assault. The author then carried out a matching of the news to the anonymous hospital data.

Karoyem and Crandall [12] carried out experiments on de-anonymization of social media users across different platforms by applying machine learning techniques. The authors collected datasets comprising photos from Flickr and tweets from Twitter. They compared different classifiers, such as decision trees, support vector machines, and Naive Bayes algorithms.

De Montjoye et al. [6] analyzed 3 months of anonymized credit card records for 1.1 million people. Their study showed that, four spatiotemporal points are enough to uniquely re-identify 90% of individuals in an anonymized dataset. The authors quantified the risk of re-identification by means of unicity.

4. RELATED WORK

Following the open data trend, different platforms emerged, providing a list of open data tools and resources, among the most widely known are: CKAN [9] a data portal software that supports individuals and organizations to make data accessible; Datahub [8], a data management platform based on CKAN that enables users to search and register published datasets; Data Portals [3], which provides a list of public administration open data portals around the world; and KDnuggets [18] aimed at providing public datasets specifically useful in the data science field. But none of them focuses on privacy issues and would assist a publisher in risk-assessment of publishing anonymized data. As a result of his extensive report on how to handle failures of anonymization, Ohm [17] recommends not to release data, even not if it is anonymized. In a similar report O'Hara discusses the issues of privacy related to the United Kingdom's transparency program [16]. In contrast, he suggests to continue anonymizing and publishing sensitive data and recommends to support the publisher. He claims, therefore steps to manage and research the risks of de-anonymization is needed.

In a related research challenge, Clifton and Marks [5] emphasised inferencing as main drawback of massive data mining. They proposed access control, auditing, data fusion, and augmentation as possible solutions to the security and privacy issues of data mining. Similarly, Li and Li [14] brought the idea of using the concept of Modern Portfolio Theory from financial investments to analyse the privacy and utility tradeoff while publishing data. In their view, the privacy-utility tradeoff in microdata publishing is similar to the risk-return tradeoff in financial investment. The authors evaluated their methodology on the Adult dataset from the UCI machine learning repository and provided quantitative interpretations to the tradeoff which can guide data publishers to choose the right privacy-utility tradeoff.

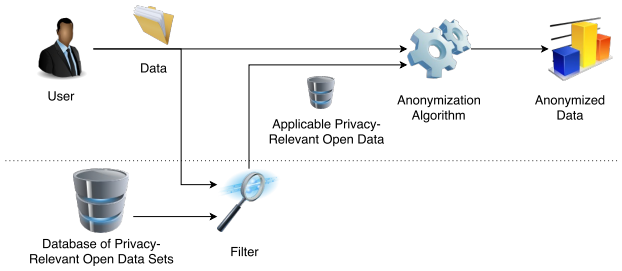

Figure 1: Mirroring Privacy-Related Open Data

5. PROPOSED APPROACH

In order to build a database with an overview of available data, organizational and technical problems have to be solved. Before we describe the steps we suggest to build this database, we discuss two different flavors: mirroring versus linking open data. We then sketch challenges and problems which may arise along with a first sketch of possible solutions in the next subsection.

5.1 Mirroring versus Linking Data

When building a database for privacy-relevant open data, the most important decision on the design of the database is, if all privacy-relevant datasets should be copied to this central database or if only a reference to the considered dataset is stored. Each implementation has its advantages and disadvantages.

Usability. We discuss usability by our primary usecase of publishing anonymized datasets while regarding open data. The usescases of testing anonymized datasets respectively de-anonymizing data follow the same pattern. If privacy-relevant open data is mirrored (Fig. 1), the user needs to reveal his data respectively only the structure of his data to the database server. On the server, privacy-relevant open data is identified and send to the user (Filter). With this additional input, an anonymization algorithm is able to produce an anonymized dataset regarding open data.

If only references to privacy-relevant open data are stored (Fig. 2), the user needs to send the structure of his data to the database server. Again, privacy-relevant open data is identified (Filter) and the corresponding references are sent to the user. The user needs to fetch those datasets by himself and is then able to apply an anonymization algorithm.

Mirroring privacy-relevant open data is more user-friendly since the user has less work and does not suffer from outages or the disappearance of open data servers.

Quality of Prediction If the user provides only the structure to the database server, both flavors do not differ much. The structure of the user's data and the structure of privacy-relevant open data may be compared. However, if the user trusts the database server and is willing to provide his data, it is also possible to additionally test for linkable data values if the privacy-relevant open data is mirrored.

Versioning. Mirroring privacy-relevant open data allows to have full control on the data, thus the data cannot be changed unnoticedly. E.g. if the regarded dataset is updated once a year, it is possible to maintain older versions to consider for the (de-)anonymization of older data. On the other hand, updates need some effort while referenced datasets are updated automatically. If privacy-relevant datasets are referenced, it is more important how they are maintained. If the datasets contain a version number and are not overwritten, but forked, it is easily possible to maintain different versions.

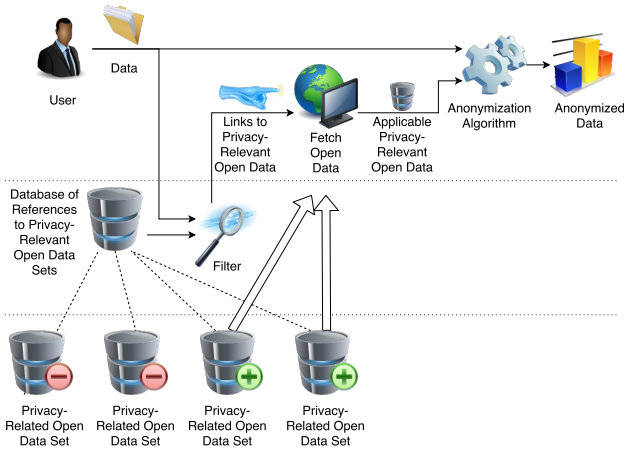

Figure 2: Referencing Privacy-Related Open Data

Storage. Mirroring privacy-relevant open data will need much more storage than only storing references.

Bandwidth. If the server mirrors the datasets and provides them to the users, a lot of upload bandwidth may be needed. On the other hand, if the service needs to work with the privacy-relevant open data a lot, storing only references may need a lot of download bandwidth since the data needs to be fetched for each computation.

5.2 Proposed Steps

Collection. First of all, privacy-relevant datasets have to be identified. In Sect. 4 we already list some tools and collections of datasets. However, none of them focuses on privacy. Thus, this is on the one hand an organizational issue which is cumbersome but not technically challenging. On the other hand, assessing the privacy aspects of a dataset is hard to be done automatically.

Linkage. After the privacy-relevant datasets are identified, it is important to test which datasets have common fields and link datasets considering common data. This way the datasets are connected and the project evolves from a collection of unlinked databases to a (partially) connected structure with data.

User-Interaction. When the database is established, an interface to the user needs to be provided. It should check if the data to be published matches available open data and provide the user with privacy-related (links to) datasets.

Anonymization. As a future step, the anonymization algorithms may be also shifted from the user to the database server. Additionally, Algorithms capable of de-anonymization may also be tried to discover if there is a privacy risk for the anonymized dataset.

5.3 Problems to Solve

For each of the steps described above, we sketch which problem(s) respectively challenge(s) arise.

Collection I. When identifying datasets relevant to privacy, it is important to notice that not all open data are relevant for privacy (e.g. datasets on sports). First, it is necessary to manually decide on the relevance of open data. If a sufficient number of datasets is judged, it may be possible to use those as a training set for machine learning algorithms and automatically decide on the relevance for future datasets.

Collection II. A version control system for privacy-related open datasets is needed since it may happen that used databases get updated, and all different versions could threaten someone's privacy. That also holds if datasets are only referenced and common problems originating from multi-sourced big data apply. Therefore, to link the datasets accordingly, also the relevant version / time frame has to be maintained.

Linkage I. Determining the context of a dataset is still an open challenge. When publishing a dataset, there is often a lack of relevant description that will indicate its associated context. E.g., a dataset regarding the population of a certain city, does not contain the specific attribute "city" within the dataset itself. Thus, if done manually, it can become a tedious task and lead to many errors caused by additional barriers (e.g., proper understanding of the associated language). In this regard, there is a need for automatic mechanisms for discovering, and determining the context of datasets. Yet, two widely adopted standards could support this tasks: the Data Catalog Vocabulary (DCAT) and the Vocabulary of Interlinked Datasets (VoID). DCAT [24] is used to describe datasets in a catalog, enabling interoperability among datasets of different provenance and increasing their discoverability. VoID [23] provides terms and patterns for describing datasets to support data discovery to cataloging and archiving of datasets, and to help users find the right data for their tasks. However, despite their benefits, a proper interpretation of the description still needs to deal with some modeling of the corresponding language, which involves additional tasks such as searching for existing vocabularies.

Linkage II. Fields in datasets do not have a unique name across different datasets. Matching those data fields in our database with those that are semantically equivalent from identified datasets requires automated recognition of data fields. A modeling strategy that deals with linguistic needs (e.g., language diversity, multilingual sources, cross-lingual linking, etc.) is required. Furthermore, identifying privacy relevant fields remains an important and open challenge. Regarding this topic privacy related ontologies came up as a basic solution, but they still need improvement considering linguistic needs.

Linkage III. Some datasets will only have partial common entries. By linking a couple of them, the resulting dataset will be quite sparse. This leads to several optimization problems. Maybe it is more promising not to make use of all associated datasets, but only of a subset of them. For n associated datasets, 2^n possible subsets exists. Which of them delivers the best results when de-anonymizing data?

User-Interaction I. Does the user need to upload all of his data, or is the structure of his data sufficient? For a first view, it is sufficient to regard only the structure of the user's data and determine if there are some intersections with the privacy-relevant open data. However, to determine if the user's data can really be linked to open data, the values need to be compared. This may be done by the user. For a more elaborated server-side analysis, the user needs to upload the anonymized dataset. In any case, after running some analysis on data from users, this data needs to be deleted properly (on the server). This also holds for all data from users which is subject of positive de-anonymization tests.

Anonymization I. Is it preferable to use a deterministic or probabilistic model ? While deterministic model may work well for small data sets, it will face efficiency and scalability problems to use it for large datasets such as the ones we intend to. Moreover, previous research by Jaro [11] has shown that probabilistic record linkage performs efficiently in large scale datasets such as public health data. Additional motivation to go for probabilistic approach is that we intend to integrate machine learning techniques in extracting privacy inferences from large scale datasets.

6. CONCLUSION

There exist frameworks that provide guidelines and tools for publishing open data (e.g., [8, 9, 10]), but most of them overlook potential privacy and security risks. Furthermore, available platforms fail to provide tools or methodologies needed to assess the level of risk of de-anonymization. Often even datasets stored in the same platform are not regarded before publishing new data. We argue that, guidelines and tools that provide mechanisms implementing de-anonymization algorithms would be of tremendous benefit.

In contrast to Ohm [17] we arrive at the conclusion that the publication of open data in general is useful, should not and cannot be prevented. Instead of trying to prevent its publication, we propose to build a database which overviews the privacy-relevant open data in a structured way. Furthermore, a tool could be built automatically linking this data to demonstrate threats on privacy. This tool and the database could serve as a source of inspiration for privacy related research. On the other hand, it could be useful for policy enhancements regarding the publication of open data.

What remains to be discussed is: if such a database exists, would it be a useful tool for publishing anonymized data or a threat to someone's privacy? As any technology, also this database could be used beyond its primary purpose, e.g. to de-anonymize already published datasets in a large scale. However, since the open data is already publicly available, this database would not leak secret information but only reduce the effort needed. This also holds for regarding open data when publishing anonymized datasets. Weighting these issues, we have to distinguish between short-term and long-term consequences. Short-term there may be privacy risks involved in creating such a database. Long-term it would show the publishers the linkability of their data to open data and improve the quality of anonymization of public datasets. If the quality of the published anonymized datasets is improved, the value of this database to adversaries decreases. Thus, we argue it is worth to build a database on open data.

7. REFERENCES

[1] Barcelona's city hall open data service. http://opendata.bcn.cat/, 2015.

[2] The CIA world factbook. https://www.cia.gov/library/publications/the-world-factbook/, 2015.

[3] Open data portals in the world. http://dataportals.org/, 2015.

[4] U.S. government's official web portal. http://www.usa.gov/, 2015.

[5] Chris Clifton and Don Marks. Security and privacy implications of data mining. In *ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, pages 15–19. Citeseer, 1996.

[6] Yves-Alexandre de Montjoye, Laura Radaelli, Vivek Kumar Singh, and Alex "Sandy" Pentland. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221):536–539, 2015.

[7] Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2006.

[8] Open Knowledge Foundation. The easy way to get, use and share data. http://datahub.io/, 2015.

[9] Open Knowledge Foundation. The open source data portal software. http://ckan.org/, 2015.

[10] Open Data Institute. What is open data? https://theodi.org/, 2015.

[11] Matthew A. Jaro. Probabilistic linkage of large public health data files. *Statistics in Medicine*, 14(5-7):491–498, 1995.

[12] Mohammed Korayem and David J Crandall. De-anonymizing users across heterogeneous social computing platforms. In *ICWSM*, 2013.

[13] Erik Lakomaa and Jan Kallberg. Open data as a foundation for innovation: The enabling effect of free public sector information for entrepreneurs. *IEEE Access*, pages 558–563, 2013.

[14] Tiancheng Li and Ninghui Li. On the tradeoff between privacy and utility in data publishing. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 517–526, New York, NY, USA, 2009. ACM.

[15] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (S&P 2008), 18-21 May 2008, Oakland, California, USA*, pages 111–125. IEEE Computer Society, 2008.

[16] Kieron O'Hara. Transparent government, not transparent citizens: A report on privacy and transparency for the cabinet office. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/61279/transparency-and-privacy-review-annex-a.pdf, September 2011.

[17] Paul Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, Vol. 57(U of Colorado Law Legal Studies Research Paper No. 9-12):1701, August 2009. Available at SSRN: http://ssrn.com/abstract=1450006.

[18] Gregory I. Piatetsky-Shapiro. Data mining, analytics, big data, and data science. http://www.kdnuggets.com/, 2015.

[19] Latanya Sweeney. Simple demographics often identify people uniquely. Technical report, Carnegie Mellon University, 2000. Data Privacy Working Paper 3.

[20] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

[21] Latanya Sweeney. Matching known patients to health records in washington state data. *CoRR*, abs/1307.1370, 2013.

[22] Peter Teufl and Günther Lackner. Rdf data analysis with activation patterns. In Klaus Tochtermann und Hermann Maurer, editor, *10th International Conference on Knowledge Management and Knowledge Technologies 1-3 September 2010, Messe Congress Graz, Austria*, Journal of Computer Science, pages 18 – 18, 2010.

[23] W3C. Describing linked datasets with the VoID vocabulary. http://www.w3.org/TR/void/, March 2011.

[24] W3C. Data catalog vocabulary (DCAT) – W3C recommendation. http://www.w3.org/TR/2014/REC-vocab-dcat-20140116/, January 2014.