PRIVA: a policy-based anonymization library for cloud and big data platform

André Ferreira, Tania Basso, Hebert Silva, Regina Moraes

[1]School of Technology – University of Campinas (UNICAMP)
Limeira – SP – Brazil

{andre.victorf,hebert.oliveiras}@gmail.com, {taniabasso,regina}@ft.unicamp.br

Abstract. *Big Data and Cloud Computing are related technologies and allow users to access data from any device. Preserving individual privacy is one of the major issues in this context, as while handling huge volumes of data it is possible that sensitive or personally identifiable information ends up disclosed. This paper presents the PRIVA, a policy-based anonymization library that aims to help providing higher reliability in cloud computing and big data platforms through privacy protection.*

Resumo. *Big data e computação em nuvem são tecnologias que se complementam e permitem que usuários acessem informações a partir de qualquer dispositivo. Preservar a privacidade dos dados neste contexto é um grande desafio pois, ao manipular uma grande quantidade de informação, é possível que dados sensíveis ou de identificação pessoal sejam divulgados sem autorização. Este artigo apresenta a PRIVA, uma biblioteca de anonimização baseada em políticas que tem como objetivo auxiliar na proteção de privacidade, provendo maior confiabilidade para plataformas de big data e computação em nuvem.*

1. Introduction

Big Data and Cloud Computing are technologies that can be combined to generate benefits and advantages for organizations. Cloud Computing enables computing resources to be provided as IT services in a pay-as-you-go fashion with high efficiency and effectiveness. Cloud-based platforms are increasingly utilized as potential hosts for Big Data, allowing the integration and analysis of large volumes of data with heterogeneous formats from different sources. Big Data analytics support the derivation of properties and correlations among data and are considered by companies a key asset to make business decisions.

Despite the numerous benefits and advantages provided by these both technologies, the analyzed data through big data analytics often include personal and sensitive information and this implies threats to privacy. So, organizations are very cautious when adopting these technologies, as there is a concern about data privacy. Usually, the reliability of these platforms are related to privacy protection, i.e., the more a platform protects the privacy of its users, clients, customers and business partners, the more credibility it gets.

One of the possible solutions to address this issue is the use of data anonymization strategies. Data Anonymization, also known as de-identification, consists of techniques that can be applied to prevent the recovery of individual information. It is mainly used to reduce the leakage of information about particular individuals while data are shared and

disclosed to public. The Anonymization process is carried out to change the data before its being disclosed. Anonymization policies, similar to privacy policies (which is a legal document that discloses the conditions under which a party can gather, use, disclose, and manage personally identifiable information), specifies how the anonymization must be performed, minimizing the risk of data re-identification. These policies may be based on privacy principles and laws, as well as data source owners specifications.

There are some anonymization tools available as, for example, ARX [Prasser and Kohlmayer 2015], SDCMicro [Templ et al. 2015], μ-ARGUS [Hundepool et al. 2005]. However, these tools perform the data anonymization based only in the knowledge of their users, requiring them to be privacy specialists and having knowledge about different data from different contexts (e.g., medical data, public transportation data, organizational and financial data, etc.). The library we present in this paper, PRIVA, is based on anonymization polices, i.e., it performs the data anonymization enforcing the rules specified in the policies. This allows improving the privacy laws compliance and data source owners privacy requirements compliance throughout the anonymization process. In addition, the fact that it is policy-based means that the proposed tool does not require its users to have such advanced knowledge in privacy, relaxing them from this task.

This paper is organized as follow: Section II presents a background and related work regarding the existent anonymization tools. Section III presents our policy-based anonymization tool. Section IV describes a case study where data from urban mobility context involving bus transportation is anonymized. Finally, Section V presents the conclusions.

2. Related Work

There are some anonymization tools available. SDCMicro [Templ et al. 2015] is a free, R-based open-source package for the generation of protected microdata for researchers and public use. This package can be used for the generation of anonymized confidential micro-data sets, i.e. for the creation of public and scientific-use files. It includes all popular disclosure risk and perturbation methods (such as global recoding, local suppression, post-randomization, micro-aggregation, adding correlated noise, shuffling and others). It also includes some risk estimation methods. The associated package sdcMicroGUI includes a graphical user interface for various methods in the sdcMicro package.

In the same software product line, SDCTAble [Meindl 2011] is a free and open source R-package to protect tabular data. It provides methods to generate instances of multidimensional, hierarchical table structures, identify primary sensitive table cells within such objects and protect primary sensitive table cells by solving the secondary cell suppression problem. This problem consists of determining which additional cells should be suppressed so that a data intruder, despite knowing the additive relationships of the tables, will not be able to estimate the sensitive cells too precisely [Massell 2001].

Similarly to SDC-family, The Argus software has the μ-ARGUS [Hundepool et al. 2005], which is a software package for the disclosure control of microdata and the τ-ARGUS [Hundepool et al. 2004] for tabular data. The package has been developed using Visual C++ language and runs under Windows versions from Windows 2000 and later. μ-ARGUS implements anonymization techniques as like

global recoding (grouping of categories), local suppression, PostRAndomisation Method (PRAM), adding noise and microaggretation. It implements a methodology for individual risk estimation based on the sampling weight. τ-ARGUS also deals with the secondary cell suppression problem [Hundepool 2004].

Other important anonymization tool for structured data is the ARX [Prasser and Kohlmayer 2015]. It supports methods for statistical disclosure control by providing: anonymization techniques such as generalization, suppression and microaggregation; privacy models such as κ-anonymity, ℓ-diversity, τ-closeness and ℓ-presence; models for analyzing reidentification risks; methods for evaluation of data utility. This tool allows anonymizing datasets with several millions of records and offers a comprehensive graphical user interface with wizards and visualizations that guide users through different aspects of the anonymization process.

Besides the previous presented anonymization tools (SDCMicro, SDCTable, μ-ARGUS, τ-ARGUS, ARX) implement lots of features, they are not policy-based and depend only of their users experience, requiring them to have a high privacy knowledge, including about privacy principles and laws. Furthermore, these tools, except ARX, are not stand-alone and work based on specifics data analytics platforms (e.g., R [Ripley 2001]). The advantage of the library proposed in this paper, that we called PRIVA, is to be based on anonymization polices. It is only needed to inform the policy and the library automatically performs the data anonymization, enforcing the rules specified in it. Furthermore, PRIVA can work stand-alone or be included in different data analytics platforms.

3. PRIVA

Anonymization, roughly speaking, is the act of removing personal identifiers from data, for example, by converting personally identifiable information into aggregated data. Anonymized data can no longer be associated with an individual [Higher Education Information Security Council (HEISC) 2015]. *Anonymization techniques* have been used to provide a balance between the beneficial use of data and the individual privacy. *Anonymization policies* can help in the anonymization process. There are regulations and guidelines to standardize the use of anonymization techniques and algorithms and the implementation of such policies are considered an important progress in organizations that want to protect their customer personal data.

PRIVA is a library developed to perform data anonymization in order to protect sensible information that is processed by data analytics algorithms and, consequently, provide higher reliability in cloud computing and big data platforms. The library is based on anonymization policies and implements the most common anonymization techniques (e.g. generalization, suppression, masking, encryption). It was developed as part of the approach presented in the work of Basso et al. [Basso et al. 2017] and the idea is that it can be integrated to that approach to perform anonymizaton in ETL process and in the two anonymization proposed phases (called *Anonymization1* for the first phase and *Anonymization2* for the second phase). PRIVA is a free and open source tool, developed in Java language, and suitable to be integrated in cloud and big data analytics platforms. Fig. 1 overviews PRIVA.

The *Privacy Library* is a Java library that has as input the data set to be anonymized and the anonymization policy that should be used to guide the anonymization

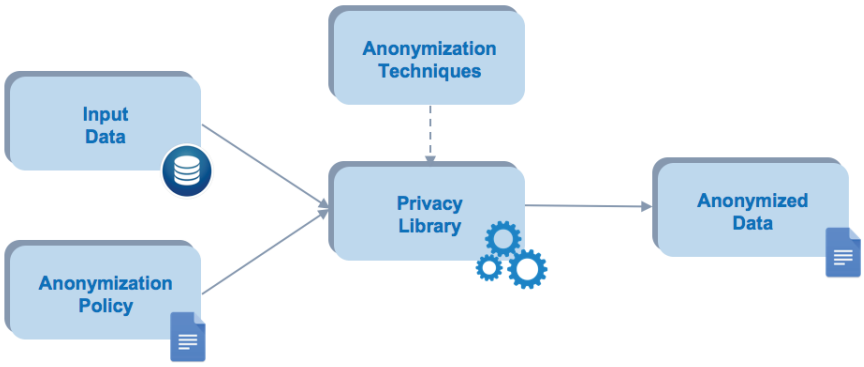

Figure 1. Library PRIVA overview

process. It applies the implemented anonymization techniques according to the policy and provides, as output, the anonymized data set.

The *Input Data* is the data from database tables and, for PRIVA, they must be loaded from files (JSON files). The fields are identified and must be anonymized according to the *Anonymization Policy*.

The *Anonymization Policy* is the guideline to the anonymization process. Basically, this policy must specify the fields related to personally identifiable information and the anonymization techniques that must be applied to each of these fields. These policies may be defined by privacy specialists, based on privacy principles and laws, as well as data source owners specifications. After defined, they can be standardized and reused by users with not so advanced privacy knowledge, facilitating the anonymization process.

The *Anonymization Techniques* refers to anonymization techniques that can be applied on data in order to protect the privacy of individual. Some of the existing and most used techniques are [Basso et al. 2016]:

- *Generalization*: attribute values are generalized to a range in order to reduce the granularity of representation. That is, the date of birth may be generalized to a range such as year of birth, so as to reduce the risk of identification;
- *Suppression*: the key attributes or the quasi-identifiers are removed completely to form the anonymized table, providing only summaries of the table data instead of individual data;
- *Encryption*: it uses cryptographic schemes to replace key-attributes, quasi-identifiers and sensitive attributes for encrypted data;
- *Perturbation (Masking)*: it consists of the replacement of the actual data values for dummy data. The idea is to randomly change the data to mask sensitive information. There are some masking techniques: (i) replacement (random replacement for similar content, but with no relation to the real data); (ii) shuffling (random replacement for data that is derived from the table column); (iii) blurring (applied to numerical data and dates. It changes the value of the data for some percentage of their random real value); (iv) reduction/nulling (replaces sensitive data with null values).

Still in Fig. 1, the *Anonymized Data* represents the resulting data set after

anonymization process. Similar to input, this output may be through JSON files.

Currently, the first release of PRIVA implements the generalization, suppression, masking (replacement) and encryption techniques. The team is working on a second release, where PRIVA will implement more techniques including the other variations of masking.

As a proof of concept, a case study was developed to show the anonymization performed by PRIVA.

4. Case Study

As a case study we performed data anonymization for the EUBra-BIGSEA project [EUBra-BIGSEA 2017]. This project address a cloud and big data platform that handles real data from transportation systems, more specifically from the Curitiba city, in Brazil. Three Big Data sources are used for the dynamic aspects of the system: the GPS location of all buses and user cards in the city (*dynamic spatial data*), data output by fine grained models of weather (*environmental data*), and historical and real-time data posted on social media associated with the city and its locations (*social network data*). This heterogeneous and large volume of data is integrated and continuously processed to detect patterns, trends and outliers in the behavior of the transportation system. The idea is to investigate speed, vehicle flux, traffic disruptions, main origin-destination routes for citizens (according to each day, time, and region), and sentiments and topics historically or recently associated with places, stress caused by traffic, landmarks, weather conditions and the effects of all of these on the perception one has from a trip in the city [EUBra-BIGSEA 2017].

This use case primarily targets two groups of end-users: citizens and urban planners. Citizen can query for the state of the route options available for a given trip. The system provides multiple route options that maximize different criteria in addition to travel time, such as likely stress, pleasantness, interestingness and liveliness of the routes (according to parameters such as day of the week, hour and location). Urban planners can obtain a descriptive view on the state of the mobility in the city as a whole, identifying its status, trends and the impacts on relevant events.

In this scenario, our goal is to provide privacy to the bus passengers. The *dynamic spatial data* is composed by (i) *user cards* data, which represents information comprising trip data per user card: the vehicle id, line code, the user card number, and the date of the trip; (ii) *vehicles and respective bus lines* data, which represents the daily itinerary, having as data the vehicle id, line code, the date time, and the latitude and longitude. With these information it is possible to perform statistical analysis in the database and track a specific user from the user card identification (e.g., it is possible to find out which bus lines were used by a specific passenger and in which localities he/she was);(iii) *personal data*, which represents the information about passengers that have an user card, as name, address, birth data, telephone, etc. When a passenger acquires his/her user card they must provide their personal information. t is important to mention that the real passenger personal information is not available in the platform and, for this case study, we generated a fake database to replace this information. Details are given in the next subsection.

4.1. The input data

Part of the input data to be anonymized in this case study are real data from public transportation from Curitiba city (Brazil). These real data refers to one working day of use of the transport system of Curitiba, in the year 2015. These data were provided for case studies in the context of the EUBra-BIGSEA project. However, the real personal information of the users, that must be associated to the bus card, were not provided, precisely for the preservation of privacy. So, to have these information, we created fictitious records.

The fictitious data was created using the *Fakenamegenerator* tool [Works 2011]. We want the fictitious user profile to be as close as possible to the actual profile. So, the fictitious data were generated respecting the proportionality of the population of Curitiba, according to the last IBGE (*Instituto Brasileiro de Geografia e Estatística* - Brazilian Institute of Geography and Statistics) census [Censo 2010], exemplified in the Age Pyramid of Figure 2.

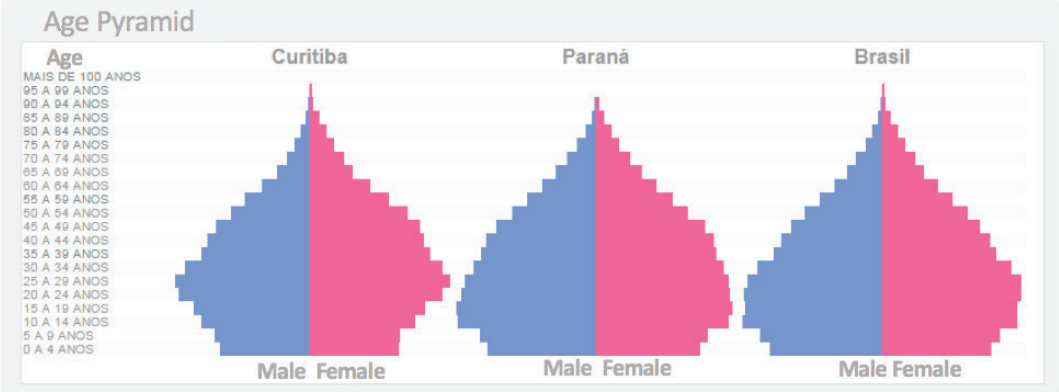

Figure 2. Curitiba's distribution of population by sex (male/female), according to age groups [Censo 2010]

In Figure 2 we can observe that the population of Curitiba city has similar proportions than Paraná State and the general population in Brazil, except for a difference in the 10 to 14 years old and 15 to 19 years old. To be as realistic as possible, we adopted the data from Curitiba.

The data correspond to one working day of use of the transport system has more than 480 thousand transactions of bus user cards. Each card is used, in average, twice a day. So, it is necessary to create more than 245 thousand user registers. We generate a table with the following fields: *Number, Gender, NameSet, Title, GivenName, MiddleInitial, Surname, StreetAddress, City, State , StateFull, ZipCode, Country, CountryFull, EmailAddress, BrowserUserAgent, TelephoneNumber, TelephoneCountryCode, Birthday, TropicalZodiac, CCType, CVV2, CCExpires, NationalID, Color, Occupation, Company, Vehicle, Domain, BloodType, Pounds, Kilograms, FeetInches, Centimeters, GUID*. The records were generated according to the data proportion from Curitiba. Then, a relationship between the tables *BusCardData*(which has data about the card, such as bus line code, bus line name, bus code, user card number, date of use) and *UserData* (which has the personal data with the fields described above) was created in order to associate each user card with a user.

4.2. The anonymization policy

We adopted the policy created for EUBra-BIGSEA [Matsunaga et al. 2017] because it is based on the existing regulations and guidelines for data anonymization found in the literature (European Data Protection Directive, GDPR (General Data Protection Regulation), PIPEDA (The Personal Information Protection and Electronic Documents Act), HIPAA (Health Insurance Portability and Accountability Act), PCI-DSS (Payment Card Industry Data Security Standard)) and built on their strength.

This policy specifies the fields related to Personally Identifiable Information and the classification of these fields into three categories, in light of the disclosure risks: (i) key attributes (attributes that uniquely identify individuals, e.g., ID, name, social security number); (ii) quasi-identifiers (attributes that can be combined with external information to expose some individuals, or to reduce uncertainty about their identities, e.g., birth date, ZIP code, position, job, blood type); (iii) sensitive attributes (attributes that contain sensitive information about individuals, e.g., salary, medical exams, credit card releases). It is a generic policy and does not include all data type that can be stored and managed by a data analytics platform, but it includes the majority of the personally identifiable information (PII) that can be held. Moreover, it can be easily extended to include other data types.

We identified, among the fields of the fictitious database, the fields that must be anonymized according to the policy. Table 1 shows the fields and respective techniques that must be applied for this case study.

Table 1. Anonymization Policy for cloud and big data-based project EUBra-BIGSEA (adapted from [Matsunaga et al. 2017])

	Field Table	Data Type	Technique	Base policy
1	Name	Key Att.	Supression	Safe Harbor method- HIPAA
2	Gender	-	Keep the data	No HIPAA guideline
3	Birth Date	Quasi-ident.	Generalization	Safe Harbor method - HIPAA
4	ID Document	Key Att.	Supression	Safe Harbor method - HIPAA
5	Address	Quasi-ident.	Supression	Safe Harbor method - HIPAA
6	City	Quasi-ident.	Supression	Safe Harbor method - HIPAA
7	State	-	Keep the data	Safe Harbor method - HIPAA
8	Country	-	Keep the data	Safe Harbor method - HIPAA
9	Postal Code	Quasi-ident.	Generalization	Safe Harbor method - HIPAA
10	Telephone Number	Quasi-ident.	Supression	Safe Harbor method - HIPAA
11	Vehicle	Quasi-ident.	Supression	Safe Harbor method - HIPAA
12	E-mail	Key Att.	Supression	Safe Harbor method - HIPAA
13	Color	-	Keep the data	No HIPAA guideline
14	CCType	-	Keep the data	Requirement 3 from PCI-DSS
15	Expiration Data	-	Keep the data	Requirement 3 from PCI-DSS
16	CVV	Sensitive Att.	Supression	Requirement 3 from PCI-DSS
17	Occupation	Quasi-ident.	Keep the data	Safe Harbor method - HIPAA
18	Company Name	Quasi-ident.	Supression	Safe Harbor method - HIPAA
19	User Card Id	Key attribute	Encryption - Hash Function	-

Besides the protection of personally identifiable information, the authors in [Matsunaga et al. 2017] present an extension of the policy to address the key attribute *user card identification*, which must be anonymized using *encryption* technique. We will also use this extension. It is represented in the latest line of Table 1, where *User Card Id*

is the number of the bus card.

Still in Table 1, the *CVV* is Card Verification Value code (CVV2 for Visa, CVC2 for Master Card and CID for AMEX), i.e., the three or four digit number located either on the front or back of a credit or debit card. *CCtype* is the Credit card type. We will use the field *Name* as the whole name, not dividing it into *GivenName, MiddleInitial, Surname*

4.3. The anonymization process

The library we propose will anonymize the input data according to the policy. It receives the both files (input data and anonymization policy) in a JSON format and generates a new JSON file with the anonymized data. Figure 3 shows the anonymization policy enforcement through PRIVA library.

Figure 3. Enforcement of anonymization policy in real bus data from Curitiba's city

In Figure 3, the *Anonymization Policy* guides the anonymization process. It describes, for each field (*FIELD_NAME*), the anonymizaton technique to be applied (*TYPE*, *DETAIL*). For exemplification, we highlighted in Figure 3 the last three fields that requires different techniques: *Company*, which requires the suppression technique (*SUP*), where the company name must be replaced by the character *; *BirthYear*, which requires a generalization (*GEN*) technique where age ranges were defined (who was born from 1900 to 1956 is classified as old; from 1957 to 1999 is adult; from 2000 to 2005 is teenager;

from 2006 to 2017 is child); *BusIdCard*, which requires the *Encryption* technique ((*ENC*), implemented as a Hash function. The other fields requires the suppression technique.

The*Database (Json) file* is the data set to be anonymized, i.e., the real bus data from Curitiba's city and the respective fictitious personal data. In this file, *busIdCard* is the field that contains the bus card identification. The other fields are the ones previously described in Section 4.1. For sake of organization, we present here just a sample of the records and highlighted the same as specified in the policy, just for comparison.

The *privacy library* applies the existing anonymization techniques and algorithms according to the policy and provides, as output, the *Anonymized Database (Json) file*, with the anonymized data set. It is possible to observe that the highlighted fields (as well as the other ones specified in the policy) are anonymized, increasing the privacy protection of these sensible information.

5. Conclusions and Future Work

In this paper we presented PRIVA, a policy-based anonymization library that helps providing higher reliability in cloud computing and big data platforms through privacy protection. Although PRIVA does not implement several resources as other available anonymization tools (e.g., data disclosure risk estimation and evaluation of data utility), the fact of being able to enforce anonymization policies makes this library a first step on (i) allowing the anonymization process be more privacy regulation compliant; (ii) standardizing and reusing anonymization policies; (iii) not requiring users to have so advanced privacy knowledge. Furthermore, PRIVA is generic, open source and can be integrated in different data analytics platforms.

The case study showed an example, as a proof of concept, of standalone use of PRIVA, enforcing a predefined anonymization policy on data from public transportation from Curitiba city. It helped us to identify errors and limitations in the library, which have been solved in order to improve it.

As future work we intend to implement more functionalities in this tool as, for example, the anonymization models (κ-anonymity, ℓ-diversity, τ-closeness). This would help minimizing the risk of data re-identification. Also, we intend to integrate the PRIVA in a data analytics platform. Once anonymization in these platforms can be performed not only on the ETL process, but also at runtime, performance evaluation shall be done and, if necessary, actions to improve the performance would be taken.

Acknowledgment

This work has been partially supported by the project **EUBra-BIGSEA** (www.eubra-bigsea.eu), funded by the Brazilian Ministry of Science, Technology and Innovation (Project 23614 - MCTI/RNP 3rd Coordinated Call) and by the European Commission under the Cooperation Programme, Horizon 2020 grant agreement no `690116`.

Also, it is supported by the project **DEVASSES** (www.devasses.eu), funded by the European Union's FP7 for research, technological development and demonstration under grant agreement no `PIRSES-GA-2013-612569`.

References

[Basso et al. 2016] Basso, T., Matsunaga, R., Moraes, R., and Antunes, N. (2016). Challenges on anonymity, privacy, and big data. In *Dependable Computing (LADC), 2016 Seventh Latin-American Symposium on*, pages 164–171. IEEE.

[Basso et al. 2017] Basso, T., Moraes, R., Antunes, N., Vieira, M., Santos, W., and Meira Jr, W. (2017). Privaaas: privacy approach for a distributed cloud-based data analytics platforms. In *International Workshop On Assured Cloud Computing And QoS Aware Big Data*, pages 1–9. IEEE.

[Censo 2010] Censo, I. (2010). Disponível em:¡ http://www. censo2010. ibge. gov. br/¿. *Acesso em*, 23.

[EUBra-BIGSEA 2017] EUBra-BIGSEA (2017). Eubra-bigsea. europe - brazil collaboration of big data scientific research through cloud-centric applications. `http://www.eubra-bigsea.eu/`.

[Higher Education Information Security Council (HEISC) 2015] Higher Education Information Security Council (HEISC) (2015). Guidelines for data de-identification or anonymization. `https://spaces.internet2.edu/display/2014infosecurityguide/Guidelines+for+Data+De-Identification+or+Anonymization`.

[Hundepool 2004] Hundepool, A. (2004). The argus-software. *Monographs of official statistics*, page 347.

[Hundepool et al. 2004] Hundepool, A., van de Wetering, A., Ramaswamy, R., de Wolf, P.-P., Giessing, S., Fischetti, M., Salazar, J.-J., Castro, J., and Lowthian, P. (2004). Userâs manual.

[Hundepool et al. 2005] Hundepool, A., Van de Wetering, A., Ramaswamy, R., Franconi, L., Capobianchi, A., DeWolf, P., Domingo-Ferrer, J., Torra, V., Brand, R., and Giessing, S. (2005). μ-argus version 4.0 software and userâs manual. *Statistics Netherlands, Voorburg NL*.

[Massell 2001] Massell, P. B. (2001). Using linear programming for cell suppression in statistical tables: Theory and practice. In *Proceedings of the Annual Meeting of the American Statistical Association*.

[Matsunaga et al. 2017] Matsunaga, R., Basso, T., Ricarte, I., and Moraes, R. (2017). Towards an ontology-based definition of data anonymization policy for cloud computing and big data. In *Manuscript submitted for publication*.

[Meindl 2011] Meindl, B. (2011). A computational framework to protect tabular data-r-package sdctable.

[Prasser and Kohlmayer 2015] Prasser, F. and Kohlmayer, F. (2015). Putting statistical disclosure control into practice: The arx data anonymization tool. In *Medical Data Privacy Handbook*, pages 111–148. Springer.

[Ripley 2001] Ripley, B. D. (2001). The r project in statistical computing. *MSOR Connections. The newsletter of the LTSN Maths, Stats & OR Network*, 1(1):23–25.

[Templ et al. 2015] Templ, M., Kowarik, A., and Meindl, B. (2015). Statistical disclosure control for micro-data using the r package sdcmicro. *Journal of Statistical Software*, 67(1):1–36.

[Works 2011] Works, C. (2011). Fakenamegenerator. `www.fakenamegenerator.com/`.