K-Anonymity technique for privacy protection: a proof of concept study

Italo Santos[1], **Emanuel Coutinho**[2], **Leonardo Moreira**[2]

[1]Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo (ICMC/USP) – São Carlos, SP - Brasil
[2]Instituto Universidade Virtual (UFC Virtual)
Universidade Federal do Ceará (UFC) – Fortaleza, CE – Brasil

italo.santos@usp.br, {emanuel,leoomoreira}@virtual.ufc.br

Abstract. *Privacy is a concept directly related to people's interest in maintaining personal space without the interference of others. In this paper, we focus on study the k-anonymity technique since many generalization algorithms are based on this privacy model. Due to this, we develop a proof of concept that uses the k-anonymity technique for data anonymization to anonymize data raw and generate a new file with anonymized data. We present the system architecture and detailed an experiment using the adult data set which has sensitive information, where each record corresponds to the personal information for a person. Finally, we summarize our work and discuss future works.*

1. Introduction

Privacy preservation has become a significant issue in many data mining applications. When a data set is released to other parties for data mining, some privacy-preserving technique is often required to reduce the possibility of identifying sensitive information about individuals [Wong et al. 2006]. According to [Cormode and Srivastava 2009], the data anonymization techniques proposed in the literature can be classified into several dimensions: **Nature of data**: Techniques have been proposed for (i) tabular data, which represents information about entities (e.g., people), their quasi-identifiers (e.g., age, gender, zip code), and their sensitive information (e.g., salary, disease); (ii) item set data, which represents transactional (or "market basket") data, associating people with the sets of items purchased in a transaction; and (iii) graph data, which represents sensitive associations between entities (e.g., people in social networks); **Anonymization approaches**: Proposed anonymization techniques use a variety of approaches, including (i) suppression, information (e.g., gender) is removed from the data; (ii) generalization, information (e.g., age) is coarsened into sets (e.g., into age ranges); (iii) perturbation, the noise is added to the data (e.g., salary); and (iv) permutation, sensitive associations between entities (e.g., purchase of medication by a person), are swapped; and **Anonymization objectives**: Various privacy goals are achieved by applying particular approaches (as above) until the resulting data has specific properties, such as (i) *k*-anonymity, each individual in the database must be indistinguishable from *k*-1 others; and (ii) other methods which aim to prevent certain inferences based on assumptions about knowledge held by an attacker.

In this paper, we focus on *k*-anonymity since many generalization algorithms are based on this privacy model. In summary, the general objective of this research is to study the *k*-anonymity technique currently used to protect sensitive data information. The specific objectives are: (i) design and develop a proof of concept (PoC) for data anonymization that uses the *k*-anonymity technique to anonymize user data; (ii) show the architecture of the PoC; and (iii) show the applicability of the PoC to a real data set.

This work is structured as follows: Section II elaborates the background of data anonymization for the notions presented and discussed in this paper. Section III briefly presents related work that implements the *k*-anonymity technique. Section IV discusses the process of development of the PoC, and the experiment is presented. Finally, the last section provides a conclusion and future works.

2. Data Anonymization

Data anonymization is used to preserve privacy over data publishing. Large private and public corporations have increasingly been charged to publish their "raw" data in electronic format, rather than providing only statistical or tabulated data. From the perspective of data dissemination of individuals, the attributes can be classified as follows [Camenisch et al. 2011]: **Identifiers:** Attributes that uniquely identify individuals (e.g. social security number, name, identity number); **Quasi-Identifiers (QI):** Attributes that can be combined with external information to expose some or all individuals, or reduce uncertainty about their identities (e.g. date of birth, ZIP code, work position, function, blood type); and **Sensitive Attributes (SAs):** Attributes that contain sensitive information about individuals (e.g. salary, medical examinations, credit card postings).

2.1. *K*-Anonymity

The concept of *k*-anonymity addresses the question of "How can a data holder release its private data with guarantees that the individual subjects of the data cannot be identified while the data remain practically useful" [Samarati 2001]. For instance, a government institution may want to release a table of data records with the names of the citizens replaced with dummy identifiers. In the experiment made in [Santos et al. 2018], a case of anonymization using government data is presented.

K-anonymization technique is a crucial component of any comprehensive solution to data privacy and has been the focus of intense research in the last few years. The *k*-anonymity model requires that any combination of QI attributes be shared by at least *k* records in an anonymous database [Samarati 2001], where *k* is a positive integer value defined by the data owner, possibly as a result of negotiations with other interested parties. A high value of *k* indicates that the anonymized bank has low disclosure risk because the probability of re-identifying a record is 1/*k*, but this does not protect the data against disclosure of attributes. Even if the attacker cannot re-identify the registry, he may discover sensitive attributes in the anonymized database.

3. Related Work

[Sweeney 2002] was the first work about k-Anonymity as an approach to sharing data in plain text without revealing private or sensitive information about individuals. The principle behind *k*-anonymity is to create *k* sets of data (equivalence classes) such that for every tuple there exist at least *k*-1 tuples that have the same QI values. The research in progress of [Fredj et al. 2014] propose guidelines as the first formalization of knowledge on anonymization algorithms for data publishers helping them to choose an algorithm given a context. The approach can be applied to select an algorithm among other anonymization techniques and even first to select a technique. In this context, our work is relevant because, we implement a PoC that implements k-anonymity technique in order to protect sensitive data, available in a web environment to facilitate use by researchers and users who are non-IT experts.

4. Proof of Concept (PoC)

Although anonymization is an essential method for privacy protection, there is a lack of tools which are both comprehensive and readily available to informatics researchers and also to non-IT experts [Prasser et al. 2014]. Graphical user interfaces (GUIs) and the option of using a wide variety of intuitive and replicable methods are needed. In this work, we develop a PoC for data anonymization, which gave rise a tool called SMDAnonymizer, that is presented in a previous work [Santos et al. 2018]. In our approach, we used the application programming interface (ARX API) to create a PoC for data anonymization, we will explain more about the development in the next section. The ARX API developed in [Prasser et al. 2014] provides a stand-alone software library with an easy-to-use public API for integration into other systems. Moreover, it implements a carefully chosen set of techniques that can handle a broad spectrum of data anonymization tasks while being efficient, intuitive, and easy to understand.

4.1. Architecture

The system architecture designed for our PoC was designed according to Model-View-Controller (MVC) architecture in web applications. In the MVC architecture pattern, the application is divided into three layers. Figure 1 displays the system architecture designed, highlighting its major components and connections.

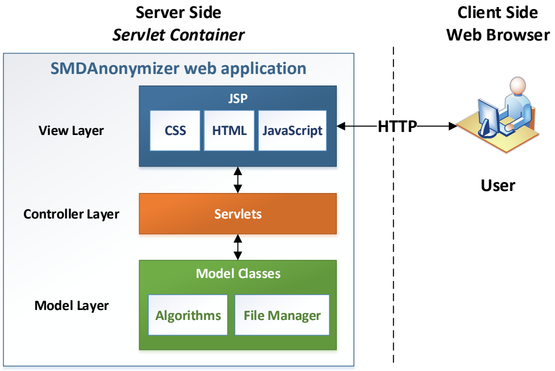


Figure 1. System architecture

The view layer has been implemented by JavaServer Pages (JSP) technology. A JSP is a technology that makes it possible to merge static content, such as HTML, with Java language instructions. In our application, JSPs were used to generate dynamic interfaces in HTML, CSS, and JavaScript. The controller layer was coded using Java Servlet technology. Servlet is a Java class that can be called through an HTTP request. Also,

Table 1. Adult Data Set Records

Sex	Age	Education	Native-country	Occupation	Salary
Male	71	Preschool	United-States	Craft-repair	<=50K
Male	18	11th	Peru	Other-service	<=50K
Female	43	Bachelors	Thailand	Machine-op-inspct	<=50K
Female	41	Doctorate	United-States	Sales	>50K
Male	33	Masters	Japan	Prof-specialty	<=50K

the Servlet can generate dynamic content, as well as JSPs, in response to a request. In SMDAnonymizer, Servlets interpret requests from JSPs, delegate processing to the model layer that executes the user request, and then formats the processing results and forwards them to the view layer.

The model layer, responsible for encapsulating the application functionalities, has two essential subcomponents: Algorithms and File Manager. The Algorithm subcomponent encapsulates the anonymization algorithms that can be used in the application. In this sense, this subcomponent is responsible for requesting, parameterizing, monitoring the execution, and retrieving the results given a data anonymization procedure. The second subcomponent, called File Manager, is responsible for manipulating files that will be anonymized by our PoC. This last subcomponent has the function of receiving the users' files, formatting, and persisting these files in the server file system to facilitate the anonymization process through the anonymization algorithms.

4.2. Settings

The proposed PoC is implemented in Java using the NetBeans[1], an integrated development environment (IDE) for coding. In our experiment, we have collected the *adult* data set provided by the University of California at Irvine (UCI) Machine Learning Repository[2]. Each record corresponds to the personal information for a person. The code of the PoC is available in GitHub[3]. The data set has the following attributes: (i) sex; (ii) age; (iii) race; (iv) marital-status; (v) education; (vi) native-country; (vii) workclass; (viii) occupation; and (ix) salary-class. In Table 1, we present an adapted example with some of the attributes. This example was used as the input data in the experiment. The PoC implements the *k*-anonymity algorithm in order to ensure that the data is anonymized. The *k*-anonymity parameter is set as standard to 2 for the results that will be presented, we intend in future versions of our application, let the user set in the interface the value of the *k*-anonymity parameter.

4.3. Process and Results

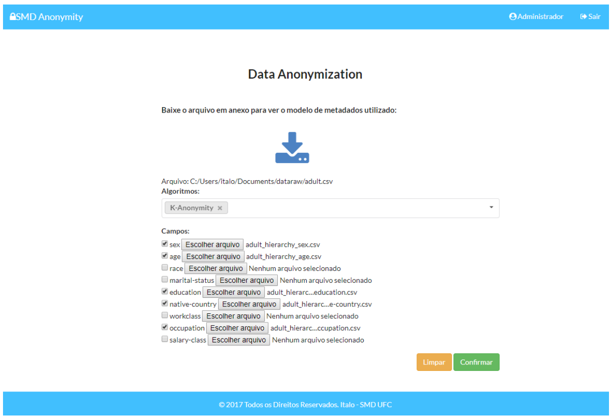
In the PoC interface is possible to download a data file example that shows to the user the type of file that it is supported (the PoC support the format *csv* short for comma separated values). This format is often used to exchange data between differently similar applications. The user has the option to use the confirm button after uploading the file, which will be anonymized. Then, the user should select the anonymization algorithm that

[1] https://netbeans.org/

[2] https://archive.ics.uci.edu/ml/datasets/adult

[3] https://github.com/italo-07/PoCTCC

will be applied to the data set. The clear button will delete the filled fields. Then the user can start uploading and selecting the anonymization algorithm again. After uploading and selecting the algorithm, the tool reads and interprets the fields referring to the columns of the data set that has been uploaded and shows in the checkbox field the columns the data set has. The user can select the fields which will be anonymized as presented in Figure 2.



SMD Anonymity

Data Anonymization

Baixe o arquivo em anexo para ver o modelo de metadados utilizado:

Arquivo: C:/Users/italo/Documents/dataraw/adult.csv
Algoritmos:
K-Anonymity ×

Campos:
- ☑ sex Escolher arquivo adult_hierarchy_sex.csv
- ☑ age Escolher arquivo adult_hierarchy_age.csv
- ☐ race Escolher arquivo Nenhum arquivo selecionado
- ☐ marital-status Escolher arquivo Nenhum arquivo selecionado
- ☑ education Escolher arquivo adult_hierarc...education.csv
- ☑ native-country Escolher arquivo adult_hierarc...e-country.csv
- ☐ workclass Escolher arquivo Nenhum arquivo selecionado
- ☑ occupation Escolher arquivo adult_hierarc...ccupation.csv
- ☐ salary-class Escolher arquivo Nenhum arquivo selecionado

© 2017 Todos os Direitos Reservados. Italo - SMD UFC

Figure 2. Step 1 — Selecting fields and uploading the anonymization hierarchies

After selecting the fields which will be anonymized, the user must upload the hierarchy. To generalize the hierarchy is created for each attribute which defines the privacy level. A hierarchy is created for QI based on the type of values these attributes hold. For instance, the hierarchy of attribute Age, Sex, Country, Occupation, and Education are shown in Tables 2, 3, 4, 5, and 6.

Table 2. "Age" attribute hierarchy

Level-0	Level-1	Level-2	Level-3	Level-4
16	15-19	10-19	0-19	*
31	30-34	30-39	20-39	*
42	40-44	40-49	40-59	*
53	50-54	50-59	40-59	*
72	70-74	70-79	60-79	*

Table 3. "Sex" attribute hierarchy

Level-0	Level-1	Level-2
Male	1	*
Female	2	*

Table 4. "Country" attribute hierarchy

Level-0	Level-1	Level-2	Level-3
United-States	USA	North America	*
Japan	JP	Asia	*
Thailand	THA	Asia	*
Peru	PER	South America	*

Table 5. "Occupation" attribute hierarchy

Level-0	Level-1	Level-2
Craft-repair	Technical	*
Sales	Nontechnical	*
Prof-specialty	Technical	*
Machine-op-inspct	Technical	*
Other-service	Other	*

Table 6. "Education" attribute hierarchy

Level-0	Level-1	Level-2	Level-3
Preschool	Primary School	Primary education	*
11th	High School	Secondary education	*
Bachelors	Undergraduate	Higher education	*
Masters	Graduate	Higher education	*
Doctorate	Graduate	Higher education	*

Therefore, after selecting fields and uploading their hierarchies, the user must confirm the operation, and then the tool exports and saves the anonymized data to a file in *csv* format. Table 7 shows the final file generated by the tool.

Table 7. Adult Data Set Records Anonymized

Sex	Age	Education	Native-country	Occupation	Salary
1	70-74	Primary School	USA	Technical	<=50K
1	15-19	High School	PER	Other	<=50K
2	40-44	Undergraduate	THA	Technical	<=50K
2	40-44	Graduate	USA	Nontechnical	>50K
1	30-34	Graduate	JP	Technical	<=50K

5. Conclusion and Future Work

In this paper, it is presented a PoC study which we applied the *k*-anonymity technique to preserve sensitive data. We have shown the system architecture and detailed an experiment that described their use to anonymize data raw and generate a new file with anonymized data, with a data set provided by UCI. Furthermore, we also identified concepts presented in the literature for data anonymization. Moreover, we presented concepts related to privacy, focusing on anonymization techniques.

As future work, we intend to further develop the PoC in order to implement other anonymization algorithms, and testing different types of data, comparing the efficiency of each implemented algorithm. Also, in future versions, we will allow the user set in the interface the value of the *k*-anonymity parameter. Moreover, we want to make a friendly interface to facilitate the user to research purposes and by users that are non-IT experts.

References

Camenisch, J., Fischer-Hübner, S., and Rannenberg, K. (2011). *Privacy and identity management for life*. Springer Science & Business Media.

Cormode, G. and Srivastava, D. (2009). Anonymized data: generation, models, usage. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 1015–1018. ACM.

Fredj, F. B., Lammari, N., and Comyn-Wattiau, I. (2014). Characterizing generalization algorithms-first guidelines for data publishers. In *KMIS 2014-International Conference on Knowledge Management and Information Sharing*, page pp.

Prasser, F., Kohlmayer, F., Lautenschläger, R., and Kuhn, K. A. (2014). Arx-a comprehensive tool for anonymizing biomedical data. In *AMIA Annual Symposium Proceedings*, volume 2014, page 984. American Medical Informatics Association.

Samarati, P. (2001). Protecting respondents identities in microdata release. *IEEE transactions on Knowledge and Data Engineering*, 13(6):1010–1027.

Santos, Í. O., Coutinho, E. F., and Moreira, L. O. (2018). Smdanonymizer: a web tool for data anonymization. In *6th International Workshop on ADVANCEs in ICT INfrastructures and Services (ADVANCE 2018)*, Santiago - Chile.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570.

Wong, R., Li, J., Fu, A. W.-C., and Wang, K. (2006). K-anonymity: An enhanced k-anonymity model for privacy preserving data publishing. In *KDD*.