Data Analysis Systems in IoE Environments for Managing Privacy and Data Protection: Pseudonymity, De-Anonymization and the Right to Be Forgotten

[1]Merugu Anand Kumar, [2]Dr. S. Gowri

[1]Research Scholar, Faculty of Computer Science and Engineering, Sathyabama Institute of Science and Technology (Deemed to be University), Jeppiaar Nagar, Rajeev Gandhi salai, Chennai -600119

[2]Professor, Department of Computer Science and Technology, Sathyabama Institute of Science and Technology (Deemed to be University), Jeppiaar Nagar, Rajeev Gandhi salai, Chennai -600119

Email -ID : [1]Meruguanand502@gmail.com, [2]gowri.it.sathyabama@gmail.com

Abstract

One of the most pressing concerns surrounding Big Data is protecting individuals' privacy, as processing massive amounts of data might lead to the exposure of private information. Actually, re-identification via privacy attacks is still possible, even with anonymised data. In order to protect large data analytics systems from re-identification risks, this article lays forth a methodology for anonymization. You may employ anonymization methods and models at two phases of this framework, which is based on anonymization policies: during the ETL process and before exporting the statistical findings of data analytics. The second step is to assess the likelihood of data re-identification and, if needed, raise the anonymity level. Although this paper presents a general framework, Ophidia was used as a case study to demonstrate how it was implemented. To ensure the re-identification procedure was successful, privacy assaults were conducted. The findings are encouraging, demonstrating a minimal likelihood of re-identification in two separate cases..

Keywords: Privacy, Anonymization, Data Analytics, Big Data.

1. Introduction:

Various institutions, including businesses and governments, have amassed vast quantities of personal data on people in recent decades. Their primary objective is to sift through this mountain of data in search of actionable insights that may improve performance in areas like public health, sales, and cost management. This kind of study is often executed using platforms that specialize in big data analytics. Concerns about privacy arise from the fact that these platforms may allow the disclosure of sensitive information to unethical parties. There must be safeguards in place to ensure that data analytics platforms do not disclose personally identifiable information in violation of privacy regulations. The European Union's plan to improve and standardize data protection for all persons is called the General Data Protection Regulation (GDPR). It was adopted and is set to be enforced in May 2018. Data anonymization is a popular method for protecting privacy in big data analytics. For data to be utilized or disseminated in a manner that does not allow for the identification of critical information, it must first be anonymized. Data usefulness might be negatively impacted, leading to inaccurate

findings and conclusions when mining, making the selection and application of these approaches a challenging process. The essential trade-off between data usefulness and anonymized data is the continued elusive pursuit of the ideal balance between privacy and utility demands. For large data analytics platforms, we provide a privacy protection approach in this study that involves re-identification risk and anonymization. This framework upgrades the one in [4], which aimed to introduce the anonymity-related privacy issues with massive data. The compromise between protecting users' privacy and keeping the data these platforms process valuable is taken into account in our enhanced architecture. The tool takes anonymization rules as input and enables the application of anonymization models and approaches in two phases: first, during the Extract, Transform, and Load (ETL) process, and second, prior to exporting the statistical findings of data analytics. We covered the components that will be employed in each step, with an emphasis on the second one. This stage assesses the danger of data re-identification and, depending on pre-established risk thresholds, raises the anonymity level to decrease this risk. To enable Ophidia, a cloud-based big data analytics platform for scientific data analysis and mining, to conduct data mining in a manner that complies with privacy and anonymity regulations, we incorporated the framework's components within. This work details one implementation in Ophidia, although the framework is flexible enough to be easily translated to other systems. With this solution, data scientists and privacy analysts can do things like: avoid privacy violations caused by attacks; perform anonymization in accordance with privacy laws and regulations and the guidelines of data source owners; and define the better balance between anonymity and data utility for different scenarios.

2. Related Work

In the literature, you may find discussions of privacy-preserving frameworks for big data analytics systems. One offered a privacy-preserving approach to e-Government frameworks that would use hashing methods to convert identifiers into digital data while protecting citizens' personal information. Using anonymization in granular access control for data analytics, another framework established an authorization approach with an emphasis on large data privacy. Another tackled problem with anonymization in common large data situations. While these works did touch on similar topics, they did not set out to analyse and implement the components discussed here. Our framework outshines these earlier attempts in three key respects: first, it integrates with the big data analytics platform rather than merely interacting with it; second, it positions the Data Utility/Re-identification Risk component as the last step in the analytics process; and third, it uses policies to guide the anonymization processes across all components, not just the ETL process. By making this change, we can be confident that boosting anonymity during ETL won't make the data less useful. Statistical Disclosure Control and Query Anonymization are both superseded by this component, which streamlines their respective functions. It successfully lessens the possibility of data re-identification, notwithstanding searches. The next part provides a detailed presentation of the framework together with its component implementations.

3. The Re- Identification Risk Based Anonymization Framework

In Figure 1, we can see the large data anonymization architecture that is based on the risk of re-identification. Data streams with massive amounts of information are also considered external data sources, as are databases that are neither relational or non-relational. The ETL

modules handle the processing of this data by combining information from several sources into a single target database. A consolidated database is represented by the Data Sources, and predictive databases are generated from the data in the Data Sources by the Derived Data Sources. The goal of data mining and analytics algorithms in a big data setting is to find useful insights in massive datasets by accessing these consolidated data sources. At two critical points where the data must pass, PRIVAaaS Anonymization and PRIVAaaS Re-identification Risk are introduced. One is during the ETL process, and the other is before the data analytics platform gives the data for display by outside users. These components can be easily modified to multiple platforms when implemented as a service, which solves the problem of interoperability. In what follows, we'll get into the specifics of how these parts work. Both parts of the anonymization process rely on policies that aim to protect users' identities. The data fields that need to be anonymized and the methods to be used are defined in these rules. Privacy regulations, rules, and the needs of data source owners (the people supplying the outside data used in analytics) are common factors that influence these decisions. Like modern Intrusion Detection Systems (IDS), the Privacy Violation Detection component keeps an eye out for and assesses any activity that might point to data exposure indicating possible privacy breaches. Despite its inclusion in the framework, this article will not touch on this aspect.

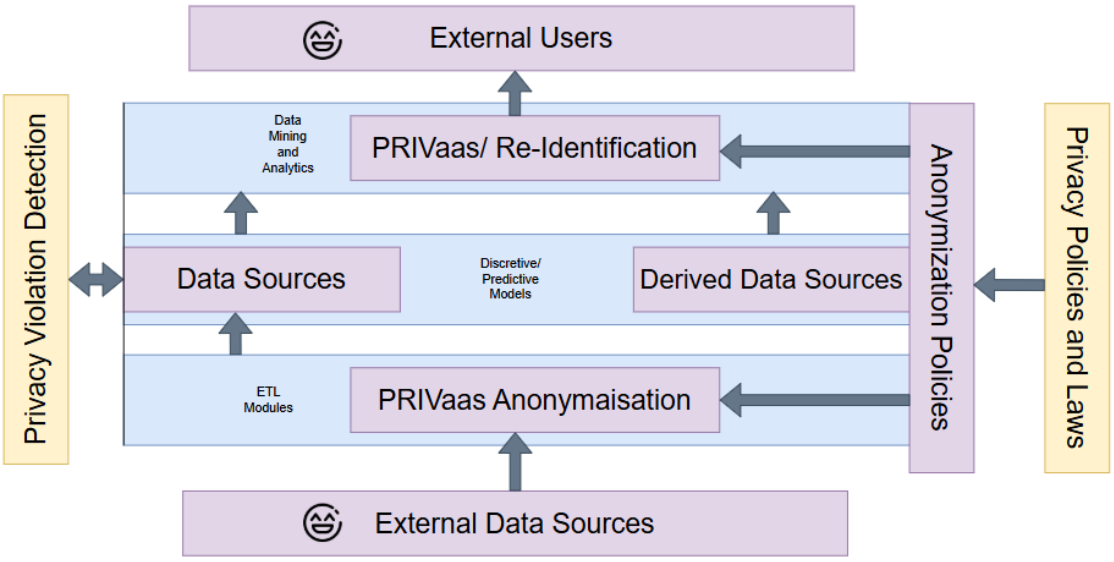


Fig 1. Anonymization Framework

With PRIVAaaS Anonymization and PRIVAaaS Re-identification Risk, the PRIVAaaS architecture is complete. Applying anonymization methods including generalization, suppression, masking, and encryption according to preset parameters, PRIVAaaS Anonymization is an open-source application designed for large data and cloud computing environments. In accordance with rules drawn from recognized standards such as PIPEDA, GDPR, PCI-DSS, and HIPAA, and particular data owner characteristics, it generates an anonymised dataset as an output. The PRIVAaaS Re-identification Risk component, on the other hand, is concerned with protecting users' privacy against data re-identification after it has left the platform, particularly in the event of privacy threats. This part determines if datasets are at danger of re-identification as a consequence of data mining and, if so, changes the

anonymity level according to predetermined levels. Identifiers or key attributes (e.g., ID, name, social security number), quasi-identifiers (e.g., birth date, ZIP code, job), sensitive attributes (e.g., salary, medical records), and non-sensitive attributes are the four types of data fields that the component takes into account when applying its anonymization policy (see Figure 2). On the basis of these categorizations, anonymization procedures are implemented, which include suppressing identifiers and sensitive fields and generalizing quasi-identifiers. For the purpose of determining whether the risk level associated with the dataset is acceptable, the anonymization policy additionally defines a re-identification risk threshold.

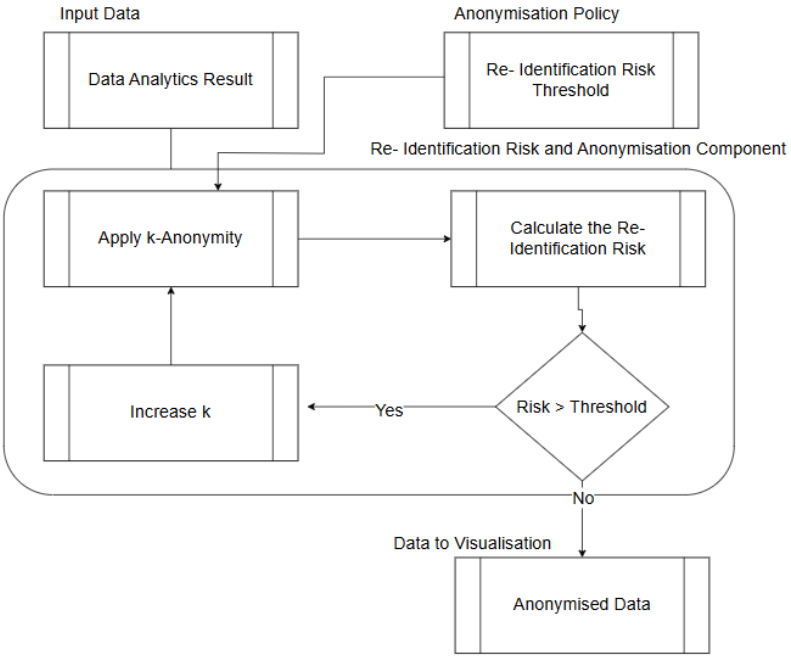


Fig 2. Re- Identification Model Component Overview

A dataset is said to have an overall risk of 90% if, for instance, each record has a 90% chance of re-identification. According to studies, research data is best served by thresholds ranging from 1% to 5%. Lower thresholds limit data usefulness, therefore selecting the barrier necessitates balancing re-identification risk with it. Data owners and privacy analysts may use this approach to implement anonymization efficiently, meeting their goals while reducing the trade-off. By identifying quasi-identifiers in the input data and applying k-anonymity, the re-identification risk component ensures that each combination of quasi-identifiers occurs in at least k entries in an anonymity table. Disclosure risk is mitigated by increasing the k-value. After implementing k-anonymity, the ARX tool and the prosecutor model profile are used to determine the re-identification risk, with a starting value of 2 representing the least value and the maximum possible danger. If the computed risk is higher than the policy threshold, the iterative process of applying k-anonymity continues until the risk is either met or falls below the threshold, after which k is increased. Ensuring compliance with privacy rules while keeping analytical value, the anonymized data is made accessible for external display after the threshold is achieved.

4. Case Study

In this section, we present a case study demonstrating the integration of the framework's components into Ophidia, a big data analytics platform. First, we introduce the platform and describe the interaction between its privacy and data analytics components during data processing. Next, we outline the tests conducted using linkage knowledge attacks to evaluate the effectiveness of the data anonymization results.

Ophidia is a research framework designed to address big data challenges in eScience by providing an environment for analysing and managing multi-dimensional heterogeneous datasets. It supports parallel and sequential operations and includes a wide range of primitives for tasks such as arithmetic functions, time-series aggregation, statistical computations, linear regression, and interpolation. However, Ophidia lacks built-in data privacy protection. Within the EUBra-BIGSEA project, Ophidia has been used to analyse data related to urban mobility, which often includes sensitive personal information requiring stringent privacy measures. To address this, the anonymization framework described earlier was integrated into the Ophidia platform.

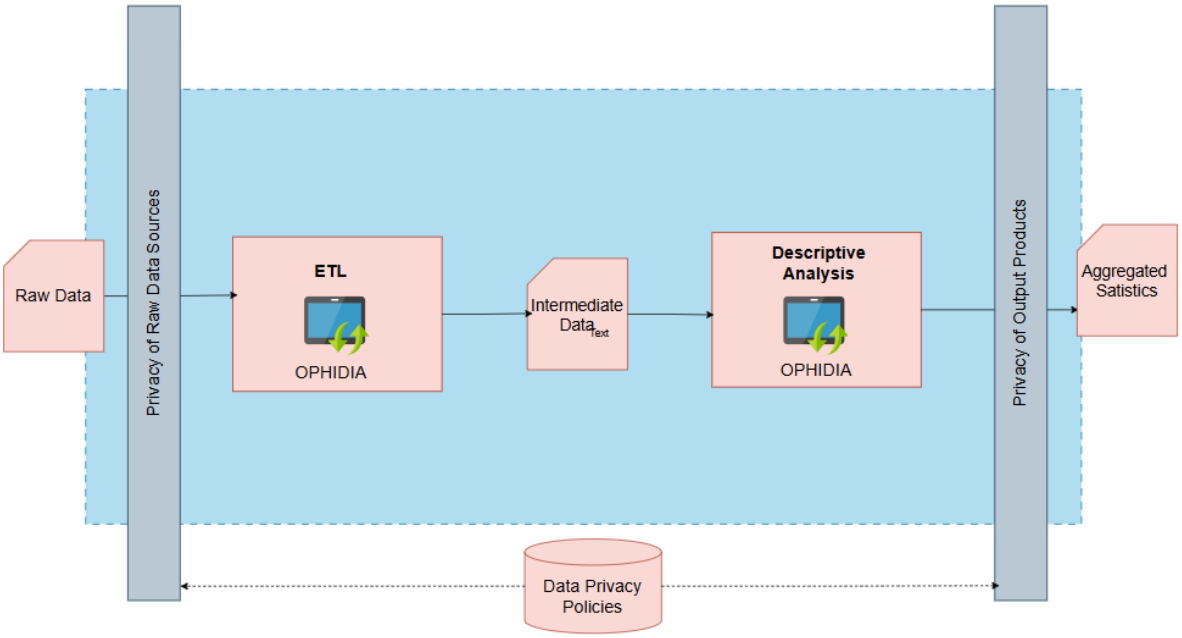


Fig 3. Privacy and Data Analytics Interaction

In order to discover statistical patterns in bus utilization using data from the Brazilian city of Curitiba, this case study was carried out within the framework of the City Administration Dashboard use case, which was created as part of the EUBra-BIGSEA project. In order to get insights while protecting user privacy, the data, which included thousands of individual bus passenger records, was aggregated and analysed. In order to achieve anonymization, components of data analytics and privacy interact during processing, as shown in Figure 3. A combination procedure is carried out in the first step of the process, which involves applying PRIVAaaS Anonymization to the raw input data. To maximize data usefulness while safeguarding the most sensitive fields, this approach uses a "AND" operation over all policy fields to ensure that only fields specified for anonymization are anonymised. As an example, the two privacy regulations in question mandated the use of an encryption method (a hash

function) to de-identify bus card identifying fields, which essentially remove any personal connections to people. The data is then subjected to ETL operations inside Ophidia after the initial anonymization. The intermediate outputs include descriptive analytics, such as statistics aggregated at different levels (e.g., bus line or user aggregation) and time frequencies (daily, weekly, monthly). Despite its usefulness, aggregated data may still include elements that may be used as quasi-identifiers, such gender and date of birth. Using the PRIVAaaS Re-identification Risk component, a supplementary anonymization step is conducted before data is sent for dashboard viewing. This second step ensures total privacy protection by increasing anonymity levels before the data leaves the Ophidia platform. Initially, less rigorous anonymization was implemented to retain analytical usability.

The second stage of anonymization, called Privacy of Output Products in Figure 3, examines the dataset for re-identification risks and, if needed, increases the level of anonymity according to a risk threshold specified in the policy file. This is accomplished by applying the k-anonymity algorithm, as mentioned earlier. Here, we see how the anonymization model (k-anonymity) used suppression and generalization approaches to mask the bus riders' identities, including their gender and date of birth, within a risk threshold range of 1% to 5%. In the first stage, known as the ETL process, there was no information loss because, according to the conjunction process, only the bus card identifier field needed to be anonymized using encryption techniques. However, if this additional anonymization process had been applied earlier, the data would have lost much of its analytical capacity. In order to maintain traceability for data analytics, the bus card values were updated using a hash function. With no data loss and a 100% re-identification risk to begin with, the primary objective of this initial step was to anonymize the most sensitive information using the conjunction method in order to minimize data loss. The k-anonymity method achieved k=2 for the 1% to 5% threshold, leading to a data loss of around 25.5%. Figure 2 shows that at lower thresholds (0.5 and 0.1%), a larger value of k was required (here, k=301), leading to complete data loss but much enhanced anonymity. With a decrease from 0.003048 to 0.000322, the re-identification risk improved by 946 percent.

Specifically, the k-anonymity technique had no effect on the risk (which remained at 0.00304878) and no further data anonymization could lower the re-identification risk for this dataset when the threshold was adjusted between 5% and 0.5%. In order to extract aggregate information, the program has been tested on three months' worth of bus card data from Curitiba, which is over 3.3GB and contains 19 million entries. To expedite the initial anonymization process, the PRIVAaaS Anonymization component was run simultaneously on several input files in addition to the ETL processes, due to the high volume of input data. At the same time, with a risk threshold of 5%, the Re-identification Risk-based component was applied to a single descriptive analytics output file. The size of this output file, which had nearly 3 million entries, was about 113MB. Preliminary findings show that the two anonymization steps impose low overhead in terms of execution time, accounting for just around 4% of the entire execution time. However, this study does not include a benchmark of the application. To further assess the effect on performance, however, more experiments are in the works.

In order to determine how well the data anonymization worked in Ophidia, studies based on linkage attacks were conducted. This kind of attack allows the attacker to find the record in the database that relates to a certain person by using supplementary information about that person. An adversary may, for instance, cross-reference a public voter list. To supplement the primary data, the tests made use of hypothetical citizen records supplied by Ferreira et al. This record

set was constructed in proportion to Curitiba's population based on the most recent Brazilian census, ensuring that it was as representative of actual citizens' records as could be. This set of information is called the "citizen sample." It contains quasi-identifier variables like gender and date of birth. A linkage attack seeks to either identify a specific human in the anonymized dataset by using the fields in the citizen sample or to find a specific citizen in the anonymized dataset by using the fields in the citizen sample. The goal of these two schools of thought is to de-anonymize the data by re-identifying persons via the identification of unique records. The citizen sample has 250,000 records, whereas the anonymized sample generated by Ophidia's analytics comprises 3,096 records. More than 253,000 assaults were generated as a consequence of the tests being repeated for every record in both the anonymized data collection and the citizen sample. Approximately 770 million searches were conducted in both datasets, since each assault entailed scanning through all entries in the target dataset. Table I shows the outcomes of the assaults carried out from the first point of view.

Table 1. Results of linkage attacks from anonymized table

Possible Matches	Occurrences	De- Anonymisation Probability
6690	207	0.0150%
7236	218	0.0137%
24139	103	0.0041%
21755	176	0.0036%
89120	955	0.0011%
95157	1239	0.0010%

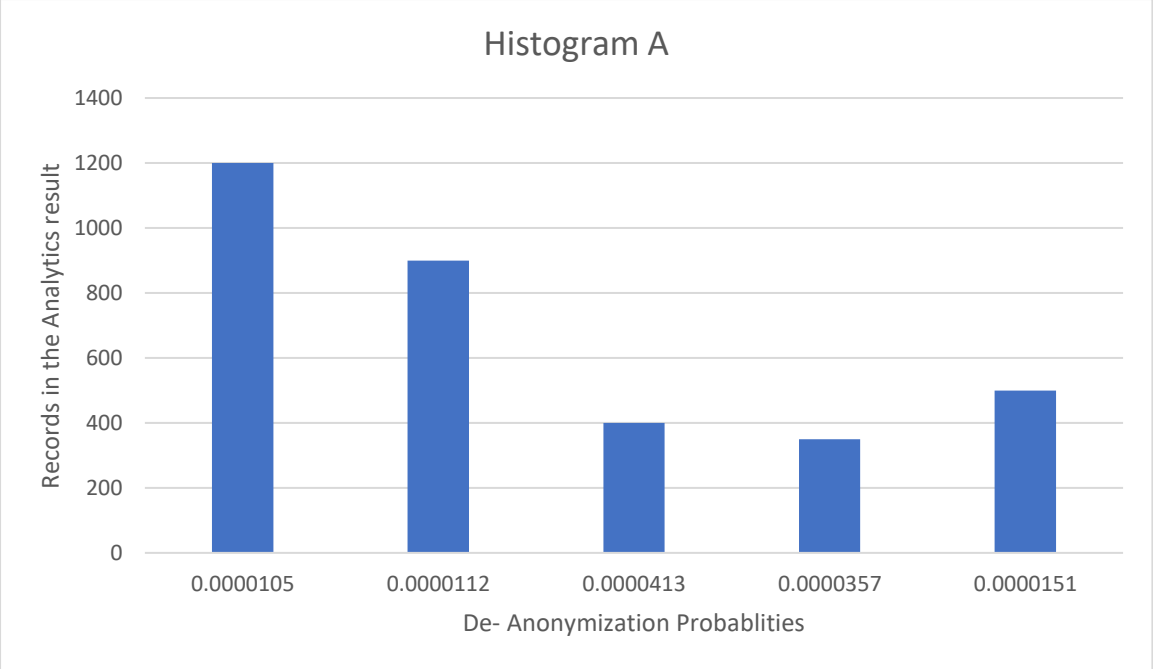
If we look at it from one angle, the attacker doesn't know who we are, and we can just count the number of times the anonymized data matches the citizen sample. A unique record from the anonymized table may potentially discover several matches in the citizen sample; this is shown in Table 1's "Possible Matches" column. In the "Occurrences" column, you can see how frequently each match possibility occurred by looking at the records when the number of match possibilities was the same. The first line of the table shows the number of potential matches for a record, which is 6,609. The second line shows the number of possible matches, which is 7,263. The subsequent lines continue in the same manner. These matches may be used to re-identify a unique record, as shown in the "De-anonymization Probability" column. The maximum risk in this experiment was 6,609 possible matches, which means that there was a roughly 0.015 percent chance of identifying a unique person from this group. In contrast, the safest option had the fewest possible matches (95,175 records) and the lowest chance of re-identification (about 0.001%). There are five separate categories of risk that emerged from the percentiles (P0 to P99) used for further analysis. There was a standard deviation of 0.005059 and over 80% of the computed probabilities were less than 0.00004134, according to this study. The statistical findings show that the maximum re-identification risk in this scenario was around 0.51% (i.e., the highest probability plus the standard deviation), which is far lower than the frequently acknowledged danger threshold of up to 5% according to the literature. These results show that the anonymization methods used in this research successfully decreased the likelihood of de-anonymization. Regarding the second point of view, the assaults mimic a situation where the assailant has more comprehensive details about the people, allowing for easier re-identification of those whose data are part of the anonymized outcome. When an

attacker has access to more context, they have a better chance of being able to re-identify people in the anonymised data, as shown in Table II.

Table 2. Results of linkage attacks from citizen sample

Possible Matches	Occurrences	De- Anonymisation Probability
1406	6651	0.0685%
1085	7237	0.0921%
176	89140	0.5988%
103	95137	0.7690%

Table II shows that the attacks yielded a maximum of 130 records as Possible Matches, which means that there was a high risk of around 0.769% in identifying a unique person in this collection of returned data. With a maximum of 1,460 records matching, the lowest risk of re-identification was achieved, with a probability of around 0.0685%. Just as in the previous experiment, we determined three different categories of risk by calculating percentiles (P0 to P99). Figure 4's Histogram B shows that the standard deviation of the computed probabilities was 0.003042, and almost 60% of those probabilities were less than 0.005989. The literature often cites a risk threshold of up to 5%, thus even with the greatest variability (over 1%), it was still below that level. After looking at these numbers, it was obvious that the data samples evaluated had low de-anonymization probability. Even with the citizen sample that is needed to de-anonymize the public transportation records from the Ophidia resultant dataset, it would be difficult for an attacker trying to target a single person to gather the exact information they need. The PRIVAaaS re-identification Risk component, which prevents the release of unique data and so reduces the likelihood of re-identification, was largely responsible for this degree of protection. This protective impact was further enhanced by the aggregated nature of the Ophidia platform's findings and the PRIVAaaS anonymization process, which together ensure that individual identities will remain low-risk even in the face of possible assaults.



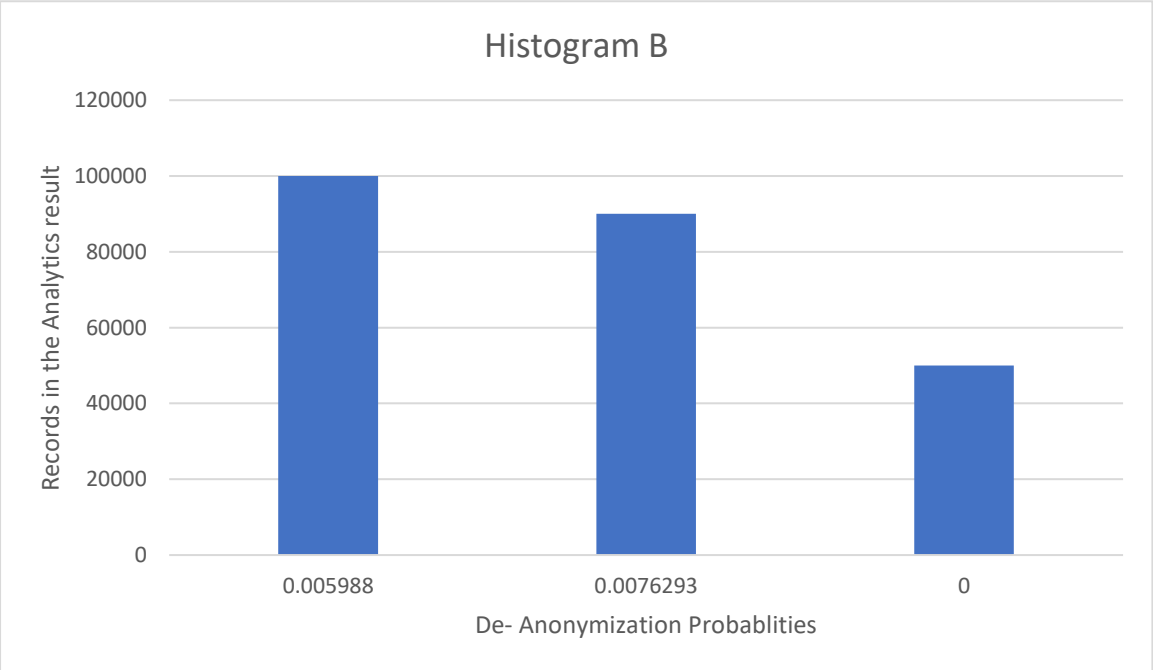


Fig 4. De-anonymization probabilities histogram

5. Insights Gained

Our investigation uncovered some intriguing conclusions about data usefulness (data loss), re-identification risk, and anonymization. One thing that has been seen is that big data analytics systems do not have anonymization options built in. There was little work required to incorporate PRIVAaaS into the Ophidia platform. The framework's components are modular and can be used with any computing application. We have started integrating it into LEMONADE, a web-based data analytics platform that lets users build ETL and machine learning workflows with a drag-and-drop interface. We are certain that these components can

be easily integrated into other data analytics systems, with the possibility of some required changes. The second takeaway is that there should have been two phases of anonymization to prevent data loss while analysing. Data loss mitigation was the primary goal of the first step of anonymization; subsequent stages strengthened privacy protections before data was made publicly available for display. By taking this tack, we were able to control the data usefulness vs. anonymization trade-off at various points. There was no chance of de-anonymization, even in the face of linking assaults, according to results from the second stage of anonymization. There was still a 1% chance of re-identification even if the worst-case scenario happened and the attacker got their hands on all the personal information from the sample of citizens. A higher or lower degree of anonymity may be achieved by adjusting these numbers according to the risk threshold, which can be set according to the demands. While more thorough performance testing is required, early results indicate that the method is viable for use in large-scale data analytics applications with no effect on platform execution times.

6. Conclusion

In this research, we detailed a framework for big data analytics platforms that uses re-identification risk to anonymize user data. We focused on the design and implementation of its primary components, which include calculating re-identification risk and applying k-anonymity. Ophidia is a data analytics platform for handling multi-dimensional heterogeneous data sets; it also included a case study that described the steps to integrate these components. The likelihood of de-anonymization, or the identifying of people inside the anonymised data collection, was assessed via the use of linkage attacks. The findings proved that the suggested solutions and integration method worked, demonstrating that the steps of anonymization effectively enhanced individual privacy protection with no effect on data usefulness. better study will expand the tests to better verify the technique utilizing additional data sets from diverse situations and deploying new forms of assaults, in addition to the present benchmark. In addition, to further improve data privacy preservation, the studies will combine various anonymization models including l-diversity and t-closeness**.**

References

1. Podda, Emanuela. "Big data analysis systems in IoE environments for managing privacy and data protection: pseudonymity, de-anonymization and the right to be forgotten." (2023).

2. Bolognini, Luca, and Camilla Bistolfi. "Pseudonymization and impacts of Big (personal/anonymous) Data processing in the transition from the Directive 95/46/EC to the new EU General Data Protection Regulation." *Computer law & security review* 33, no. 2 (2017): 171-181.

3. Tudor, Valentin, Magnus Almgren, and Marina Papatriantafilou. "The influence of dataset characteristics on privacy preserving methods in the advanced metering infrastructure." *Computers & Security* 76 (2018): 178-196.

4. Barta, Gergő. "Challenges in the compliance with the General Data Protection Regulation: anonymization of personally identifiable information and related information

security concerns." *Knowledge–economy–society: business, finance and technology as protection and support for society* (2018): 115-121.

5. Esayas, Samson. "The role of anonymisation and pseudonymisation under the EU data privacy rules: beyond the 'all or nothing'approach." *European Journal of Law and Technology* 6, no. 2 (2015).

6. Pikulík, Tomáš, and Peter Štarchoň. "GDPR compliant methods of data protection." In *6th SWS International Scientific Conferences on social sciences 2019 Conference proceedings*, pp. 561-572. 2019.

7. Bagnato, Alessandra, Paulo Silva, Ala Sarah Alaqra, and Orhan Ermis. "Workshop on privacy challenges in public and private organizations." *Privacy and Identity Management. Data for Better Living: AI and Privacy: 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2. 2 International Summer School, Windisch, Switzerland, August 19–23, 2019, Revised Selected Papers 14* (2020): 82-89.

8. Davis, John S., and Osonde A. Osoba. "Privacy Preservation in the Age of Big Data: A Survey." (2016).

9. Casini, Paola. "Data protection in the European Union institutions from an information management perspective." In *Recordkeeping in International Organizations*, pp. 28-58. Routledge, 2020.

10. Fischer-Hübner, Simone, Marit Hansen, Jaap-Henk Hoepman, and Meiko Jensen. "Privacy-Enhancing Technologies and Anonymisation in Light of GDPR and Machine Learning." In *IFIP International Summer School on Privacy and Identity Management*, pp. 11-20. Cham: Springer Nature Switzerland, 2022.

11. Tahir, Hassane, and Patrick Brézillon. "Contextualization of Personal Data Discovery and Anonymization Tools." In *Intelligent Sustainable Systems: Selected Papers of WorldS4 2021, Volume 1*, pp. 277-285. Springer Singapore, 2022.

12. Tsakalakis, Niko, Sophie Stalla-Bourdillon, and Kieron O'hara. "Data protection by design for cross-border electronic identification: Does the eIDAS interoperability framework need to be modernised?." *Privacy and Identity Management. Fairness, Accountability, and Transparency in the Age of Big Data: 13th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2. 2 International Summer School, Vienna, Austria, August 20-24, 2018, Revised Selected Papers 13* (2019): 255-274.

13. Guarino, Alessandro. "What now? Data Retention Scenarios After the ECJ Ruling." In *ISSE 2014 Securing Electronic Business Processes: Highlights of the Information Security Solutions Europe 2014 Conference*, pp. 249-255. Wiesbaden: Springer Fachmedien Wiesbaden, 2014.

14. Sartore, Federico. "Big data: privacy and intellectual property in a comparative perspective." (2016).

15. Aleksieva-Petrova, Adelina, Ivaylo Chenchev, and Milen Petrov. "Three-Layer Model for Learner Data Anonymization." In *INTED2020 Proceedings*, pp. 6244-6252. IATED, 2020.