ARANI: Uma Abordagem Baseada em Linha de Experimento para Preservação de Privacidade em *Data Lakes*[*]

Thiago Jordão[1], **Marcos Bedo**[1], **Daniel de Oliveira**[1]

[1]Instituto de Computação – Universidade Federal Fluminense (UFF)

thiagojordao@id.uff.br, {marcosbedo,danielcmo}@ic.uff.br

Resumo. *Os Data Lakes armazenam grandes volumes de dados heterogêneos, incluindo informações sensíveis. Garantir conformidade com regulamentações como a LGPD exige o uso de técnicas de anonimização. Técnicas aplicadas de forma isolada, como k-Anonimato ou Privacidade Diferencial, podem ser insuficientes. A combinação dessas técnicas em fluxos configuráveis é, portanto, essencial. Linhas de Experimento permitem estruturar e instanciar esses fluxos de forma flexível. Este artigo propõe a* ARANI*, uma abordagem baseada em Linha de Experimento que permite definir, executar e avaliar fluxos de anonimização com suporte a múltiplas técnicas.*

Abstract. *Data Lakes store large volumes of heterogeneous data, including sensitive information. Ensuring compliance with regulations such as the LGPD requires the use of anonymization techniques. Techniques applied in isolation, such as k-Anonymity or Differential Privacy, may be insufficient. Therefore, combining these techniques in configurable flows is essential. Experiment Lines enable the flexible structuring and instantiation of these flows. This paper proposes* ARANI*, an Experiment Line-based approach that allows the definition, execution, and evaluation of anonymization flows with support for multiple techniques.*

1. Introdução

Os *Data Lakes* são ambientes para armazenamento e análise de grandes volumes de dados heterogêneos [Nargesian et al. 2019]. Entre os principais benefícios estão: *(i)* a flexibilidade para lidar com dados estruturados, semiestruturados e não estruturados; *(ii)* o tempo reduzido de implantação; e *(iii)* o suporte a arquiteturas distribuídas para armazenamento e consultas. Um componente essencial nos *Data Lakes* é o *Data Publisher* [Bauer et al. 2022], responsável pela ingestão, organização e disponibilização dos dados. Ele é o componente que coleta dados de fontes externas, realiza pré-processamento e enriquecimento, e os armazena no *Data Lake*.

Nos últimos anos temos visto um aumento na preocupação com a privacidade de dados em *Data Lakes*, pois muitos arquivos contêm dados sensíveis que permitem identificar indivíduos e, consequentemente, estão sujeitos a LGPD. Segundo [Oreščanin et al. 2024], o *Data Publisher* é um ponto crítico, que exige atenção e demanda a utilização de estratégias para mitigar a ameaça de publicação de dados sensíveis. Uma das principais estratégias é a Anonimização de Dados, que transforma os dados para dificultar re-identificações, seja pela supressão de atributos ou adição de ruído em atributos. Contudo, a aplicação de tais técnicas de forma isolada pode ser insuficiente. [Domingo-Ferrer and Torra 2005] demonstram que métodos tradicionais de anonimização não preservam adequadamente semântica e precisão quando aplicados simultaneamente a dados numéricos e categóricos. Além disso, a combinação de diferentes técnicas

[*]Os autores gostariam de agradecer pelo apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Código de Financiamento 001, do CNPq e da FAPERJ.

pode mitigar ataques de diferenciação, por exemplo. Nesse cenário, a aplicação de técnicas de Privacidade Diferencial [Dwork et al. 2006] e k-Anonimato [Sweeney 2002] pode limitar o vazamento de informações.

Entretanto, esse fluxo pode ser complexo de ser modelado, e pode ser visualizado como uma Linha de Experimento [Ogasawara et al. 2009], que representa uma família de fluxos com funcionalidades comuns e variações conforme os requisitos dos usuários. No contexto deste artigo, essa linha de experimento permite instanciar fluxos de anonimização no *Data Lake*, adaptados às características dos dados. Nesse cenário, propomos a `ARANI` ("proteção" em Tupi-Guarani), uma abordagem que instancia fluxos de anonimização com base em uma Linha de Experimento. A `ARANI` executa consultas em bancos de dados relacionais e aplica fluxos de anonimização parametrizados. A `ARANI` permite a adição de novas técnicas de anonimização e fornece métricas para estimar riscos de privacidade e utilidade dos dados. A abordagem foi avaliada com um *dataset* de dados do Censo dos EUA, apresentando resultados promissores.

2. Referencial Teórico e Trabalhos Relacionados

2.1. Técnicas de Anonimização de Dados

Existem diversas técnicas de anonimização de dados, cada uma com características e limitações distintas, sendo mais ou menos adequadas conforme o tipo e o contexto dos dados. Entre as abordagens mais simples, destaca-se a supressão, que pode ser total, com a omissão completa de um atributo sensível, ou parcial, substituindo partes do dado por caracteres genéricos. Apesar de simples de implementar, a supressão pode comprometer a utilidade dos dados. Em contraste, outras técnicas introduzem perturbações nos dados, mantendo padrões estatísticos globais. Entre elas estão o *swapping*, que troca valores entre tuplas distintas, e a microagregação, que agrupa tuplas em *clusters* e substitui os valores por médias ou medianas.

Há também abordagens formais como o k-Anonimato [Sweeney 2002] e a l-Diversidade [Machanavajjhala et al. 2007]. O k-Anonimato garante que cada tupla em um *dataset* seja indistinguível de pelo menos $(k-1)$ outras quanto aos atributos *quasi*-identificadores [Zigomitros et al. 2020][1]. A l-diversidade estende esse conceito ao exigir que cada grupo equivalente contenha pelo menos l valores distintos para o atributo sensível. Destacam-se ainda técnicas que introduzem ruído estatístico, como a Privacidade Diferencial [Dwork et al. 2006], que assegura que a inclusão ou exclusão de um indivíduo não altere significativamente os resultados de análises. Isso é feito por meio da adição de ruído aleatório controlado às respostas das consultas, regulado por um parâmetro ϵ, que define o nível de privacidade. Valores menores que ϵ indicam maior proteção, com possível perda de precisão. A técnica de Privacidade Diferencial é implementada por meio de mecanismos de distribuição estatística. Entre os principais estão os mecanismos de Laplace e Gaussiano. O primeiro é indicado para consultas numéricas e adiciona ruído proporcional à sensibilidade da consulta e inversamente proporcional a ϵ. Já o mecanismo Gaussiano utiliza ruído da distribuição Gaussiana e atende ao modelo relaxado de (ϵ, δ)-privacidade, permitindo uma probabilidade de falha na garantia de privacidade e oferecendo maior flexibilidade.

2.2. Linhas de Experimento

Uma Linha de Experimento [Ogasawara et al. 2009] é uma abstração usada para representar as transformações de dados em experimentos. Essa abordagem se baseia em três conceitos: *(i)* va-

[1]Atributos cujas características são publicamente conhecidas e que, isoladamente, não identificam um indivíduo, mas que, quando combinados, podem reduzir o conjunto de possíveis indivíduos, aumentando o risco de reidentificação.

riabilidade, *(ii)* opcionalidade e *(iii)* derivação. A *Variabilidade* se refere às diferentes formas de implementar uma mesma transformação. Neste artigo, quando uma anonimização pode ser realizada por duas ou mais técnicas, ela é tratada como um *ponto de variação*. Caso contrário, é uma transformação *invariante*. A Figura 1 apresenta um exemplo de linha de experimento para anonimização de dados onde as transformações "Generalização" e "Adição de Ruído" representam pontos de variação, enquanto "Classificação de Atributos" e "Avaliação da Privacidade" são operações invariantes.

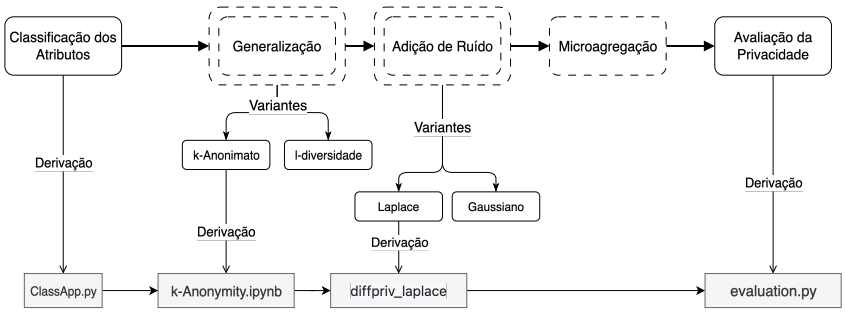

Figura 1. Um exemplo de uma Linha de Experimento com cinco transformações.

A *Opcionalidade* indica se uma transformação pode ou não estar presente nos fluxos de anonimização. No exemplo da Figura 1, as transformações "Generalização", "Adição de Ruído" e "Microagregação" são opcionais, enquanto "Classificação dos Atributos" e "Avaliação da Privacidade" são obrigatórias. A *Derivação* é o processo de gerar fluxos de anonimização executáveis (*scripts* Python), a partir de uma descrição abstrata do experimento, com base nas escolhas do usuário sobre pontos de variação e transformações opcionais. Na Figura 1, os retângulos cinza representam os *scripts* que serão executados para realizar de fato a anonimização.

2.3. Trabalhos Relacionados

A publicação de dados com preservação da privacidade é um tema relativamente recente, mas fundamental com a expansão dos *Data Lakes* e das legislações de dados pessoais. [Machado and Amora 2021] analisam o impacto dessas leis nos SGBDs e identificam componentes para garantir conformidade. [Oreščanin et al. 2024] propõem o PEDAL-MM (***Pe**rsonal identifiable information **Da**ta **L**ake **M**etadata **M**odel*), voltado ao tratamento de dados pessoais em *Data Lakes*. No entanto, o modelo apenas registra metadados, sem abordar a tarefa de publicação dos dados. [Deshpande 2021] propõe a Sypse, que usa pseudo-anonimização e particionamento, distribuindo identificadores entre partições de dados para garantir privacidade.

A Aircloak [Francis et al. 2018] anonimiza resultados de consultas `SQL`, mas não exporta dados para *Data Lakes*. Ferramentas como o DATPROF e *Oracle Data Masking* aplicam transformações simples, mas carecem de técnicas de privacidade robustas, permanecendo vulneráveis a reidentificação. Ferramentas como Amnesia [Terrovitis et al. 2012], SECRETA [Poulis et al. 2014] e ARX [Prasser et al. 2020] suportam publicação com privacidade. A Amnesia não permite configurar atributos *quasi*-identificadores e sensíveis, enquanto que SECRETA e ARX permitem essa definição e oferecem várias técnicas e métricas de utilidade. A ARX ainda oferece suporte à privacidade diferencial. A ferramenta PRIVAS [Miguel et al. 2019] permite configurar níveis de privacidade, suportando k-Anonimato, l-diversidade e privacidade diferencial. No entanto, a ARX não permite empilhar técnicas sobre o mesmo *dataset* nem adicionar novos algoritmos.

3. A Abordagem ARANI

A arquitetura da ARANI é ilustrada na Figura 2 e é composta por nove módulos principais: *(i)* Anotação dos Atributos, *(ii)* Banco de Metadados de Privacidade, *(iii)* Registro de Algoritmos, *(iv)* Catálogo de Algoritmos, *(v)* Registro da Linha de Experimento, *(vi)* Instanciação do Fluxo de Anonimização, *(vii)* Execução do Fluxo de Anonimização, *(viii)* Publicação dos Dados no *Data Lake* e *(ix)* Banco de Dados de Proveniência.

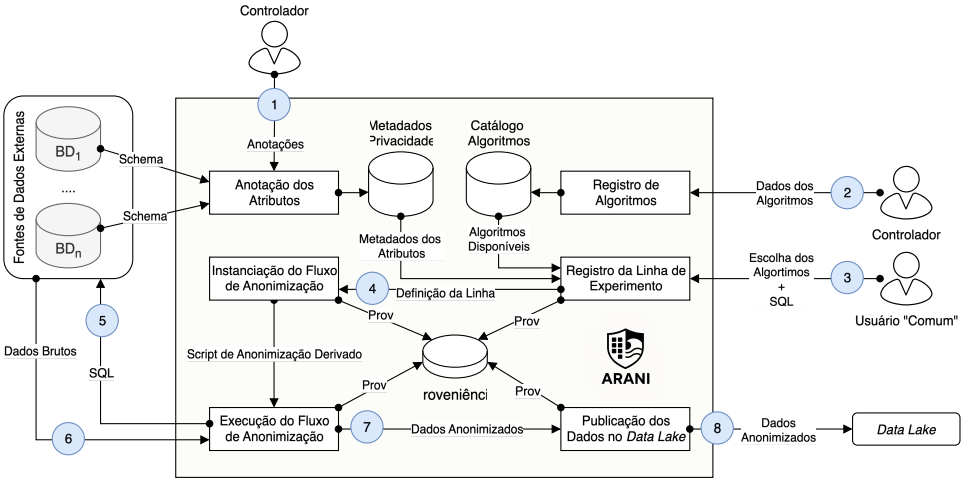

Figura 2. Arquitetura da ARANI.

A ARANI define dois perfis de usuário: *(i)* o Controlador, responsável pela configuração do ambiente, e *(ii)* o Usuário "Comum", encarregado da criação e instanciação das linhas de experimento. O processo tem início com o Controlador, que realiza a *Anotação dos Atributos* das tabelas de cada *schema* das fontes de dados externas (passo ① na Figura 2). Em conformidade com a legislação vigente, o Controlador atribui um nível de privacidade a cada atributo, seguindo a classificação definida por [Barros et al. 2024]. Essas anotações são armazenadas no *Banco de Metadados de Privacidade*. Além de anotar os atributos, o usuário Controlador deve cadastrar os algoritmos que podem ser utilizados na ARANI no *Registro de Algoritmos*. Para cada implementação de técnica de anonimização, é necessário realizar o *upload* do código, dos parâmetros e da linha de comando utilizada para executar o código (passo ② na Figura 2). Todas essas informações são armazenadas no *Catálogo de Algoritmos.* Assume-se que o usuário responsável por realizar o *upload* do código seja também o proprietário dos dados a serem anonimizados. Assim, eventuais falhas ou comportamentos maliciosos nos *scripts* resultariam em prejuízo ao próprio usuário, o que desincentiva práticas inadequadas.

Com os algoritmos registrados e os atributos anotados, o usuário pode criar sua linha de experimento no módulo de *Registro da Linha de Experimento* (passo ③ na Figura 2). Nesse módulo, são definidas as etapas de anonimização no fluxo, os comandos SQL para extração de dados das fontes e os algoritmos que podem ser utilizados em cada etapa. Todas as definições da linha de experimento são armazenadas no *Banco de Dados de Proveniência*.A linha só precisa ser configurada uma única vez e pode ser usada múltiplas vezes *a posteriori* para instanciar e executar os fluxos de anonimização. Após a criação da linha de experimento, o usuário inicia a *Instanciação do Fluxo de Anonimização* (passo ④ na Figura 2), selecionando os *scripts* a serem executados em cada etapa. Esses *scripts* são instanciados a partir do Ca*tálogo de Algoritmo*s e executados no módulo de *Execução do Fluxo de Anonimização*, que processa os comandos SQL (passo ⑤ na Figura 2) e aplica o fluxo de anonimização sobre os dados brutos recebidos

(passo ⑥ na Figura 2). Os dados anonimizados são então enviados ao módulo de *Publicação dos Dados no Data Lake* (passo ⑦ na Figura 2), onde são armazenados de acordo com o padrão definido (passo ⑧ na Figura 2), como por exemplo, arquivo .csv, visão materializada ou arquivo Parquet. Todos os metadados da execução são registrados no *Banco de Dados de Proveniência*.

Ao final do processo de anonimização e carga dos dados no *Data Lake*, a ARANI calcula métricas de risco e utilidade dos dados anonimizados. O nível de privacidade é quantificado com base nos índices de risco propostos por [Giomi et al. 2023] e considera três métricas: *(i)* RID: risco de identificação direta de um indivíduo; *(ii)* RA: risco de associação com dados externos; *(iii)* RIF: risco de inferência de atributos sensíveis. A utilidade dos dados (U) é mensurada para avaliar o quanto das características analíticas ou estatísticas dos dados originais foi preservado após a anonimização. Para consolidar os aspectos de privacidade e utilidade em uma métrica única, definiu-se a Função Objetivo de Anonimização (FOA). A utilização da FOA busca quantificar o nível de eficiência do experimento. Essa função penaliza casos em que o nível de privacidade ou a utilidade são excessivamente comprometidos, seguindo a formulação: $FOA = \theta(1 - (avg(RID, RA, RIF)) + \lambda U$, onde: *(i)* RID, RA e RIF representam os índices de risco calculados com $\{RA, RI, RF\} \in [0, 1]$, *(ii)* U é o valor da utilidade dos dados anonimizados, e *(iii)* $\theta, \lambda \in \mathbb{R}^+$ são parâmetros de ajuste, fornecidos pelo Controlador, que ponderam a importância do risco e utilidade no custo total. O código-fonte da ARANI se encontra disponível em https://github.com/UFFeScience/arani.

4. Avaliação Experimental da ARANI

Esta seção apresenta a avaliação experimental da ARANI. Utilizamos o *dataset Adult* [Becker and Kohavi 1996], baseado no Censo dos EUA de 1994, cujo objetivo é prever se um indivíduo recebe mais de US$ 50K por ano, com base em dados demográficos e socioeconômicos. O *dataset* original contém 48.842 tuplas e 15 atributos e, após a remoção de tuplas duplicadas, restaram 45.175 tuplas. As tuplas foram então carregadas no PostgreSQL, e a ARANI foi integrada a esse ambiente. O objetivo é modelarmos uma linha de experimento para anonimização, conforme ilustrado na Figura 1.

Os atributos do *dataset* foram classificados conforme definido por [Barros et al. 2024]. Em seguida, foi definida a consulta SQL para extração dos dados que retornava todos os atributos do *dataset*, e foram criadas quatro variantes com diferentes fatores de seletividade, retornando 25%, 50%, 75% e 100% dos dados. A avaliação foi realizada em três rodadas e, em cada rodada, o mesmo fluxo de anonimização foi instanciado, fixando-se as técnicas utilizadas (privacidade diferencial sobre o resultado do k-Anonimato) e variando-se os atributos *quasi*-identificadores e sensíveis, além da quantidade de tuplas a serem anonimizadas (de acordo com o fator de seletividade). Para a privacidade-diferencial, adotou-se os seguintes parâmetros: $\varepsilon = 0{,}5$ e $\delta = 1 \times 10^{-5}$. O ruído foi calibrado utilizando a distribuição Gaussiana com desvio padrão $\sigma = \frac{\sqrt{2\ln(1{,}25/\delta)} \cdot \Delta f}{\varepsilon}$, com $\Delta f = \frac{\max(x) - \min(x)}{n}$. Para o k-Anonimato, foram considerados os valores k = 1.000 e k = 4.000, refletindo cenários com diferentes exigências de privacidade.

Os conjuntos de atributos *quasi*-identificadores utilizados na técnica k-Anonimato foram: c_1 = {*"age"*, *"workclass"*, *"education"*}, $c_2 = c_1 \cup$ {*"fnlwgt"*}, $c_3 = c_2 \cup$ {*"marital-status"*, *"occupation"*, *"relationship"*, *"race"*, *"gender"*, *"native-country"*}. Além disso, foi utilizado o atributo "*"capital-gain*" como atributo sensível na técnica (ϵ, δ)-privacidade diferencial em todos os conjuntos. É importante mencionar que a ordem de aplicação das técnicas de anonimização pode influenciar os resultados obtidos, mas por limitações de espaço, optamos por explorar exclusivamente a configuração em que o k-Anonimato é aplicado antes da privacidade

diferencial. As métricas de risco de privacidade e utilidade dos dados anonimizados (Tabela 1) são calculadas após a execução de cada fluxo de anonimização. Esses indicadores permitem quantificar a eficiência do fluxo instanciado apresentando valores entre 0 (baixa eficiência) e 1 (alta eficiência) para as métricas, calculados por meio da *Anonymeter* [Giomi et al. 2023].

Tabela 1. Resultados agrupados por k, conjunto de atributos, fator de seletividade f_s, $P =$ risco de privacidade, $U =$ utilidade e $T =$ tempo de processamento.

k		1.000			4.000		
Conjunto	$f_s(\%)$	P	U	T	P	U	T
c_1	25	0.143	0.875	35.3	0.198	0.870	23.6
	50	0.108	0.880	66.7	0.102	0.875	51.8
	75	0.104	0.881	104.5	0.112	0.878	77.8
	100	0.113	0.884	139.2	0.088	0.879	106.0
c_2	25	0.012	0.846	48.9	0.016	0.833	40.5
	50	0.009	0.849	109.7	0.014	0.846	73.6
	75	0.012	0.854	154.7	0.012	0.848	110.0
	100	0.011	0.854	191.0	0.006	0.846	129.7
c_3	25	0.007	0.776	49.0	0.011	0.720	34.0
	50	0.005	0.788	103.5	0.011	0.779	76.2
	75	0.012	0.797	143.2	0.014	0.767	113.3
	100	0.010	0.795	179.0	0.008	0.775	141.5

Para cada valor de k, o tempo de processamento (em segundos) intraconjunto aumenta proporcionalmente ao fator de seletividade f_s. Também se observa que esse tempo diminui à medida que k cresce. Esse comportamento foi consistente em todos os conjuntos de atributos. Na análise interconjunto, o tempo de processamento de c_2 em relação a c_1 aumentou, em média, 47% e 44% para $k = \{1.000, 4.000\}$, respectivamente. A comparação entre c_2 e c_3 mostra uma redução média de 4% para $k = 1.000$ e nenhuma diferença relevante para $k = 4.000$. Esses resultados indicam que o k-Anonimato é mais sensível à distribuição estatística dos atributos *quasi*-identificadores, visto que a inclusão do atributo "*fnlwgt*" em c_2 teve mais impacto que a adição dos demais atributos em c_3. Na comparação entre c_1 e c_2, quando "*fnlwgt*" passa a ser *quasi*-identificador, observa-se queda no risco de privacidade com a utilidade constante. Isso reforça que esse atributo é determinante no aumento do risco, devido à sua natureza identificável. Na comparação entre c_2 e c_3, o risco permanece estável, sugerindo que a principal redução já ocorreu com a anonimização de "*fnlwgt*". A utilidade dos dados permaneceu constante na análise inter-conjuntos para todos os valores de k, assim como na comparação entre c_1 e c_2. Já entre c_2 e c_3, observa-se queda de utilidade para todos os valores de k. Isso indica que a utilidade é mais sensível à proporção de dados anonimizados do que às características estatísticas dos atributos marcados como *quasi*-identificadores ou sensíveis.

5. Conclusão

Este artigo apresentou a `ARANI`, uma abordagem baseada em Linhas de Experimento para a preservação de privacidade em dados exportados para *Data Lakes*. A `ARANI` oferece uma solução flexível para publicação segura de dados pessoais sensíveis ao integrar diferentes técnicas de anonimização em fluxos de anonimização configuráveis e extensíveis. A arquitetura modular proposta permite o reúso de componentes, facilita a configuração de novos fluxos e incorpora métricas que auxiliam na avaliação dos *trade-offs* entre privacidade e utilidade. Os experimentos realizados com o *dataset Adult* demonstraram a viabilidade da abordagem, indicando como a escolha criteriosa de atributos *quasi*-identificadores impacta nos níveis de risco e utilidade dos dados. Como trabalhos futuros, pretende-se expandir o catálogo de algoritmos, incorporar mecanismos de aprendizado de máquina para sugestão de configurações ideais e realizar estudos de caso em ambientes reais de produção.

Referências

Barros, P. V. d. S. et al. (2024). Incorporando os requisitos e as restrições da lgpd ao projeto de banco de dados. In *SBBD'24*, pages 341–353. SBC.

Bauer, D. et al. (2022). Revisiting data lakes: the metadata lake. In *Middleware'22*, page 8–14, New York, NY, USA.

Becker, B. and Kohavi, R. (1996). Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Deshpande, A. (2021). Sypse: privacy-first data management through pseudonymization and partitioning. In *CIDR*, pages 1–8, Chaminade, CA.

Domingo-Ferrer, J. and Torra, V. (2005). Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195–212.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *TCC 2006*, volume 3876, pages 265–284. Springer.

Francis, P., Probst-Eide, S., Obrok, P., Berneanu, C., Juric, S., and Munz, R. (2018). Diffix-birch: Extending diffix-aspen. *arXiv preprint arXiv:1806.02075*.

Giomi, M. et al. (2023). A unified framework for quantifying privacy risk in synthetic data. *Proceedings on Privacy Enhancing Technologies*, 2023(2):312–328.

Machado, J. C. and Amora, P. R. (2021). The impact of privacy regulations on db systems. *Journal of Information and Data Management*, 12(5).

Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. (2007). L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1):3–es.

Miguel, J., Pereira, M. J., Henriques, P., and Berón, M. (2019). Assuring data privacy with privas – a tool for data publishers. *IADIS International Journal on Computer Science and Information Systems*, 14(2):41–58.

Nargesian, F., Zhu, E., Miller, R. J., Pu, K. Q., and Arocena, P. C. (2019). Data lake management: Challenges and opportunities. *Proc. VLDB Endow.*, 12(12):1986–1989.

Ogasawara, E. et al. (2009). Experiment line: software reuse in scientific workflows. In *Proc. of the SSDBM 2009*, pages 264–272, Berlin. Springer.

Oreščanin, D., Hlupić, T., and Vrdoljak, B. (2024). Managing personal identifiable information in data lakes. *IEEE access*, 12:32164–32180.

Poulis, G. et al. (2014). SECRETA: A system for evaluating and comparing relational and transaction anonymization algorithms. In *EDBT'14*, pages 620–623.

Prasser, F., Eicher, J., et al. (2020). Flexible data anonymization using arx—current status and challenges ahead. *Software: Pract. and Exp.*, 50(7):1277–1304.

Sweeney, L. (2002). k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570.

Terrovitis, M., Liagouris, J., Mamoulis, N., and Skiadopoulos, S. (2012). Privacy preservation by disassociation. *arXiv preprint arXiv:1207.0135*.

Zigomitros, A., Casino, F., Solanas, A., and Patsakis, C. (2020). A survey on privacy properties for data publishing of relational data. *Ieee Access*, 8:51071–51099.