Anonimização de Streaming de Dados em DOCA

Bruno C. Leal[1], **Israel C. Vidal**[1], **Javam C. Machado**[1]

[1]Laboratório de Sistemas e Banco de Dados (LSBD)
Universidade Federal do Ceará (UFC) – Fortaleza – CE – Brasil

{bruno.leal,israel.vidal,javam.machado}@lsbd.ufc.br

Abstract. *Online scenarios are increasingly more common, providing great opportunities for data analysis. Such data usually contains sensitive information and should be anonymized to guarantee individuals' privacy. This work proposes DOCA, a differentially private approach for publishing data streaming in non-interactive scenarios using an online microaggregation strategy to obtain better utility.*

Resumo. *Cenários online são cada vez mais comuns, propiciando grandes oportunidades de análise de dados. Frequentemente, esses dados contém informações sensíveis de indivíduos e, portanto, devem ser anonimizados para garantir sua privacidade. Este trabalho propõe DOCA, uma abordagem diferencialmente privada para publicação de streaming de dados em cenários não interativos utilizando uma estratégia de microagregação online para obtenção de melhor utilidade.*

1. Introdução

A coleta de dados vem crescendo em grandes proporções nos últimos anos, tanto em volume quanto em velocidade. Esses dados apresentam valiosas oportunidades para análise, e.g. técnicas de aprendizado de máquina para classificação ou predição. Cenários *online*, tais como monitoramento e análises de ações da bolsa, análise de densidade de tráfego, *IoT* e *Smart cities*, estão se tornando cada vez mais comuns. Porém, grande parte das informações são oriundas de dados sensíveis dos indivíduos e devem ser tratadas com cuidado para garantir a privacidade dos mesmos.

Privacidade Diferencial (PD) [Dwork and Roth 2014] tem se tornado o modelo padrão de privacidade de dados em detrimento de outros modelos como k-anonimato [Sweeney 2002]. Sua aplicação se dá, principalmente, na proteção de informações estatísticas obtidas por consultas interativas. Neste cenário, o resultado das consultas em formato agregado, como por exemplo histogramas, limita as possibilidades de análises. Para oferecer maior capacidade de análise, a publicação do dado em cenário não interativo, numa versão anonimizada do dado original, é desejável [Soria-Comas et al. 2014]. Todavia, para a publicação de dados em cenário não interativo utilizando PD, a quantidade de ruído necessária para proteger cada item do conjunto de dados pode ser tão grande que o dado anonimizado gerado pode não ter utilidade.

A privacidade de indivíduos no contexto de *streaming* de dados é, portanto, extremamente relevante e deve ser atacada. Este trabalho propõe, então, a abordagem DOCA (***D**ifferential Privacy **O**nline **C**lustering and **A**nonimization*) para garantir a Privacidade Diferencial de forma a minimizar o ruído no contexto de publicação não interativa, garantindo maior liberdade de análise para uma *streaming* de dados numéricos.

2. Trabalhos Relacionados

Para a aplicação da Privacidade Diferencial em consultas que retornam dados numéricos (e.g. *COUNT*), o Mecanismo de Laplace [Dwork and Roth 2014] é o mais amplamente utilizado e aceito. Este gera um ruído a ser adicionado à resposta real com base em dois parâmetros: *(i)* o *budget* ϵ e *(ii)* a sensibilidade Δf. O primeiro é um parâmetro de entrada para ajustar o nível de privacidade desejado. Já a sensibilidade depende do domínio da consulta e é definida como o máximo impacto possível de um indivíduo na resposta. A sensibilidade é utilizada para garantir que a presença ou ausência de um indivíduo no dado não irá afetar substancialmente a resposta.

PD é adotada, principalmente, no contexto interativo para consultas de informações estatísticas [Chen et al. 2011, Xu et al. 2013]. Entretanto, trabalhos nesse contexto limitam as possibilidades de análises e não se relacionam diretamente com o tipo de publicação do DOCA. No contexto de publicação de versões anonimizadas no formato original do dado, i.e., versões não interativas, os trabalhos [Zhang et al. 2017, Bindschaedler et al. 2017] geram tuplas sintéticas a partir de modelos construídos com base no dado original. Porém, não cobrem o contexto de *streaming* de dados.

Os trabalhos [Soria-Comas et al. 2014, Soria-Comas and Domingo-Ferrer 2017] publicam uma versão diferencialmente privada do conjunto de dados microagregado. Os autores argumentam que a perda de informação pela microagregação, que consiste em substituir cada valor de um grupo pelo centroide, é compensada pela redução do ruído necessário para proteger o dado microagregado. O DOCA utiliza a sensibilidade para geração do ruído de um grupo definida em [Soria-Comas and Domingo-Ferrer 2017] como $\Delta c = \Delta f/|C|$, onde C é o grupo a ser publicado e Δf a sensibilidade. O ruído para o grupo a ser publicado é gerado a partir do mecanismo de Laplace com média zero e escala $\Delta c/\epsilon$. Entretanto, essa estratégia é voltada para a anonimização de um conjunto de dados estático, enquanto o DOCA realiza a anonimização sobre um fluxo de dados contínuo.

3. DOCA - *Differential Privacy Online Clustering and Anonymization*

DOCA é uma abordagem de publicação de *streaming* de dados numéricos univariados que aplica a Privacidade Diferencial em um contexto não interativo, ao mesmo tempo que diminui a sensibilidade na adição de ruído para, ao publicar, manter a utilidade do dado sem comprometer a garantia de privacidade dos indivíduos.

No contexto de atuação do DOCA, é necessário considerar características específicas de *streaming* de dados que impactam fortemente no processo de anonimização por Privacidade Diferencial, a saber [Silva et al. 2013]: (1) o dado é potencialmente ilimitado, o que impede sua representação por inteiro em memória; (2) o algoritmo de microagregação deve ser *online*; (3) cada registro de entrada está sujeito a uma restrição de tempo de processamento (*delay constraint*) entre sua entrada e saída.

A Figura 1 ilustra as duas etapas do DOCA: (i) agrupamento *online* e (ii) anonimização com microagregação e adição de ruído.

3.1. Agrupamento *Online*

A anonimização realizada pelo DOCA adiciona, além da perda de informação originada pela microagregação, o ruído advindo do mecanismo de Laplace [Dwork and Roth 2014]

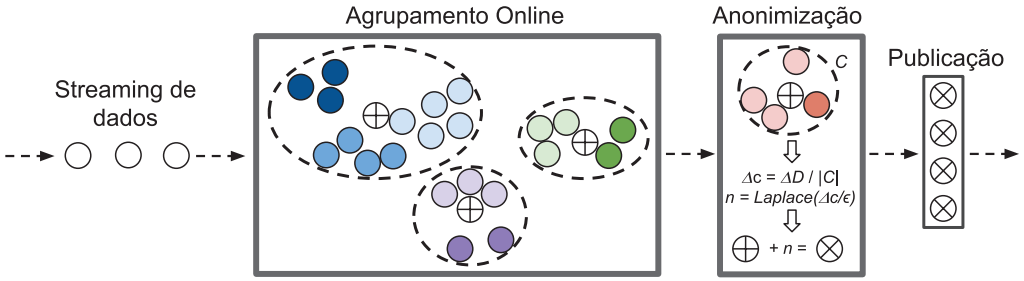

Figura 1. Uma visão geral da abordagem proposta.

para garantir a Privacidade Diferencial. Este ruído é diretamente proporcional à sensibilidade (Δf) e inversamente proporcional ao tamanho do grupo sobre o qual ele é aplicado pois a sensibilidade adotada no DOCA é $\Delta c = \Delta f/|C|$, onde C é o grupo a ser publicado.

Nós argumentamos que quanto maior o tamanho do grupo, menor quantidade de ruído se faz necessária, desde que o centroide seja representativo o suficiente para minimizar a perda de informação decorrente da microagregação. Então, para manter boa representatividade enquanto maximiza-se o tamanho dos grupos, criamos um novo algoritmo de clusterização online como uma variação da solução oferecida em [Cao et al. 2011]. Para garantir que os grupos tenham centroides representativos, utiliza-se um limiar para a máxima perda de informação de um grupo, dado por τ, que é continuamente atualizada (Algoritmo 1 linha 11) como a média da perda de informação dos últimos m grupos publicados. Então, dado que a representatividade de um grupo é garantida, esse grupo crescerá até que um de seus elementos atinja a *delay constraint*. Portanto, o agrupamento feito pelo DOCA no Algoritmo 2 nas linhas 3 a 5, onde a escolha do melhor cluster para um item (linha 4) é separado no Algoritmo 2 para melhor legibilidade, visa maximizar o tamanho dos grupos sem perder a representatividade dos centroides.

Algoritmo 1: DOCA

Entrada: $S, \Delta f, \beta, m, \epsilon$
Saída: Atualiza os clusters online C
/* S: streaming de dados */
/* Δf: sensibilidade global */
/* β: número máximo de clusters em memória */
/* m: número máximo de clusters usados para computar τ */
/* τ: média da perda de informação dos últimos m clusters publicados */
1 C = [ ], I = [ ], $\tau = 0$, $min_t = +\infty$, $max_t = -\infty$;
2 **para** *cada novo Registro* $r \in S$ **faça**
3 min_t = min(r, min_t); max_t = max(r, max_t);
4 Cluster c = BestSelection(r, C, τ, min_t, max_t);
5 adicione r em c;
6 Registro r_e = GetExpiringRecord();
7 **se** r_e != *Nulo* **então**
8 Cluster $c_e = \{c_i \in C | r_e \in c_i\}$;
9 i_{loss} = (max(c_e) - min(c_e)) / (max_t - min_t);
10 adicione i_{loss} em I;
11 τ = média dos últimos m valores $\in I$;
/* Passo 2-anonimização e publicação do cluster c_e */

Algoritmo 2: BESTSELECTION Seleciona o melhor grupo para um novo item

Entrada: r, C, τ, min_t, max_t
Saída: Cluster c_{best}

```
1  C_min = [ ], C_best = [ ], e_min = +∞;
2  para cada c ∈ C faça
3      e_i = GetEnlargment(r, c);
4      e_min = min(e_min, e_i);
5  para cada c ∈ C faça
6      se GetEnlargment(r, c) == e_min então
7          adicione c em C_min;
8          c_test = cópia de c; adicione r em c_test;
9          i_lossTest = (max(c_test) - min(c_test)) / (max_t - min_t);
10         se i_lossTest < τ então add c em C_best ;
11 se C_best é ∅ então
12     se tamanho de C < β então
13         c_new = novo Cluster; adicione c_new em C;
14         retorna c_new;
15     senão retorna um cluster de C_min com o menor tamanho ;
16 senão retorna um cluster de C_best com o menor tamanho ;
17
```

3.2. Anonimização

Essa etapa consiste em verificar se existe uma tupla que atingiu a *delay constraint* e, por isso, deve ser publicada (Algoritmo 1 linha 6). Para isso, publica-se o grupo ao qual essa tupla pertence, i.e., a tupla juntamente com as demais tuplas do mesmo grupo. Os passos de anonimização desse grupo que ocorrem após a linha 11 do Algoritmo 1 foram omitidos por questão de espaço e são descritos a seguir. Para publicação do grupo calcula-se o valor de seu centroide χ como a média dos valores. Após isso computa-se o valor do ruído η a ser inserido nesse grupo, dado por uma amostra aleatória obtida da distribuição de Laplace com média zero e escala $\Delta c/\epsilon$ (Δc definido na Subseção 3.1). Substitui-se, então, o valor de todas as tuplas por $\chi + \eta$ e, finalmente, publica-se o grupo anonimizado e remove-se o mesmo juntamente com seus elementos do conjunto de dados ainda em processamento e não publicados.

Observe que Δc não depende do tamanho do conjunto de dados, portanto pode ser diretamente aplicada no contexto de microagregação em *streaming* de dados. Além disso, como argumentado em [Soria-Comas and Domingo-Ferrer 2017], uma mesma amostra de ruído de Laplace deve ser gerada para todo o grupo. Caso contrário, se fosse gerado um ruído diferente para cada tupla, a quantidade de ruído gerada seria maior do que na abordagem tradicional, pois dessa forma teria-se, além do ruído da abordagem tradicional, a perda de informação decorrente do passo de microagregação. Perceba, também, que é satisfeita a restrição de limitação de memória, bem como a restrição de tempo de processamento até a saída, pois como o DOCA trata a chegada de uma nova tupla como uma unidade de tempo, jamais haverá mais tuplas em processamento do que o valor estabelecido para a restrição *delay constraint*.

4. Avaliação Preliminar

Para a avaliação experimental, foi simulada uma *streaming* de dados a partir do conjunto de dado BackBlaze [Backblaze 2017] no qual utilizamos o atributo *smart_9_raw*. O con-

junto de dado apresenta as seguintes características: 1.989.462 registros, 42.762 valores distintos, valor mínimo 1, valor máximo 66.413 e *skew* 0,7365. O *skew* foi medido através do Coeficiente de Assimetria dado por $\frac{1}{n*s^3}\sum_{i=1}^{n}(X_i - \bar{X})^3$, onde n é o tamanho do conjunto de dado, s é o desvio padrão e $\bar{X}$ é a média. O experimento foi repetido cinco vezes para todos os itens e, em cada execução, o item a ser processado foi selecionado aleatoriamente uma única vez.

Para computar os resultados apresentados, calculou-se a média das cinco execuções. Este critério foi adotado pois, como o algoritmo é *online*, a ordem em que os dados chegam pode ser relevante para o resultado final. O número de execuções deve ser baixo para não viciar o resultado, pois como característica do mecanismo de Laplace, a média de várias execuções tende a aproximar o resultado do valor real. Em relação aos parâmetros de entrada do DOCA foram utilizados os seguintes valores: *delay constraint* = 1000, $\beta = 50$, $m = 100$, escolhidos arbitrariamente e podem ser variados para buscar melhor utilidade; $\epsilon = 1,0$ pois é um valor aceito na literatura capaz de garantir a privacidade dos indivíduos e ainda manter bom nível de utilidade; sensibilidade $\Delta f = 99.618,0$, conforme utilizado por [Soria-Comas and Domingo-Ferrer 2017] onde $\Delta f = 1,5 * |max(conjuntodedado) - min(conjuntodedado)|$.

Para análise de utilidade comparou-se o DOCA com a estratégia ingênua de publicar os dados sem implementar o passo de microagregação *online*, simplesmente adicionando uma amostra do ruído de Laplace com média zero e escala $\Delta f/\epsilon$ a cada tupla. Avalia-se dessa forma pois, de acordo com nosso conhecimento, não há nenhum trabalho até o momento que trate da publicação diferencialmente privada de uma *streaming* de dados no contexto não interativo. Para essa avaliação utilizou-se o Erro Quadrático Médio (EQM) para representar a diferença numérica entre os valores reais e anonimizados. Os resultados são exibidos na Tabela 1.

Tabela 1. Avaliação de EQM entre DOCA e Abordagem ingênua.

EQM da abordagem ingênua	EQM do DOCA	Diminuição do EQM(%)
20.297.943.835,9597	143.064.439,6199	99,2952

Já para mostrar o quanto o dado anonimizado com o DOCA representa o dado original, mede-se o percentual de interseção entre as distribuições. A Figura 2 apresenta visualmente a comparação entre a distribuição original e a versão anonimizada através de um histograma com cem *bins* para uma das execuções. A média da área de interseção das cinco execuções obtida foi de $85,98\%$.

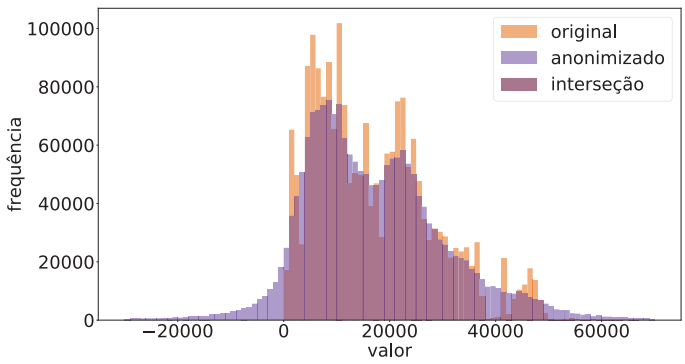

Figura 2. Distribuição original x anonimizada. Interseção = 86,01%.

5. Conclusão

Este artigo apresentou a estratégia DOCA para a anonimização de *streaming* de dados para publicação, fazendo uso de adição de ruído laplaciano nos agrupamentos desses dados. Os resultados preliminares comprovaram uma significativa redução de ruído adicionado, ao mesmo tempo que mantiveram um alto nível de privacidade, preservando ainda as características do conjunto de dados original. Várias oportunidades de trabalhos futuros são identificadas a partir dos nossos resultados. Dentre elas, podemos enumerar: (i) adaptar o DOCA para suportar *streaming* de tuplas com múltiplos atributos; (ii) definição de um modelo para ajuste dos parâmetros da fase de agrupamento *online*; (iii) suporte a atributos categóricos; e, talvez o mais importante, (iv) suporte a múltiplas tuplas de um mesmo indivíduo.

Referências

Bindschaedler, V., Shokri, R., and Gunter, C. A. (2017). Plausible deniability for privacy-preserving data synthesis. *Proc. VLDB Endow.*, 10(5):481–492.

Cao, J., Carminati, B., Ferrari, E., and Tan, K. L. (2011). Castle: Continuously anonymizing data streams. *IEEE Transactions on Dependable and Secure Computing*, 8(3):337–352.

Chen, R., Mohammed, N., Fung, B. C. M., Desai, B. C., and Xiong, L. (2011). Publishing set-valued data via differential privacy. *PVLDB*, 4(11):1087–1098.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407.

Backblaze (2017). The raw hard drive test data from 2017-01-01 to 2017-01-31. Online at `https://www.backblaze.com/b2/hard-drive-test-data.html`. acessed 2018-04-22.

Silva, J. A., Faria, E. R., Barros, R. C., Hruschka, E. R., Carvalho, A. C. P. L. F. d., and Gama, J. a. (2013). Data stream clustering: A survey. *ACM Comput. Surv.*, 46(1):13:1–13:31.

Soria-Comas, J. and Domingo-Ferrer, J. (2017). Differentially private data sets based on microaggregation and record perturbation. In *MDAI 2017, Kitakyushu, Japan, October, 2017, Proceedings*, pages 119–131.

Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D., and Martínez, S. (2014). Enhancing data utility in differential privacy via microaggregation-based k-anonymity. *The VLDB Journal*, 23(5):771–794.

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

Xu, J., Zhang, Z., Xiao, X., Yang, Y., Yu, G., and Winslett, M. (2013). Differentially private histogram publication. *The VLDB Journal*, 22(6):797–822.

Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., and Xiao, X. (2017). Privbayes: Private data release via bayesian networks. *ACM Trans. Database Syst.*, 42(4):25:1–25:41.